The Dissertation Committee for Jared Dale Fisher
certifies that this is the approved version of the following dissertation:

Balancing Model Structure and Flexibility in Forecasting Financial Time Series

Committee:

Carlos M. Carvalho, Supervisor

Jared S. Murray

Davide Pettenuzzo

Thomas S. Shively

Balancing Model Structure and Flexibility in Forecasting Financial Time Series

by

Jared Dale Fisher

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2019

Dedicated to my wife Danielle.

Acknowledgments

I would like to first thank my mentors and advisors. Carlos, you have been a tremendous advisor. Thank you for giving me opportunity to do good work and the encouragement to succeed. Davide, thank you for guiding me through my first work in econometrics. Tom, thank you for the enlightening discussions on monotonicity and time-dynamics. Jared, thank you for coaching me on both causal inference and computation. James, thank you for being an inspiring teacher. Furthermore, thank you all for your efforts supporting this body of work and my future career steps. I would be amiss if I forget to also thank Gil, Shane, and Jeff for the opportunity to start learning how to do research almost a decade ago.

I am also very grateful for my fellow PhD students and these experiences we shared together. There are many of you, and I wish you all the best in your own journey! A special thanks goes to David. Thank you for being my brother in the program and my friend in life. I look forward to the success of our current and future work together.

Finally, I thank my family. I'm ever grateful for my wife Danielle, to whom this work is dedicated. May we ever pursue knowledge, improvement, and progression together. Words befall my gratitude for your support and encouragement, through the good times and the hard. I'm grateful for the patience and cheer of our children, as they support their father's work in their own ways. I'm grateful for the support of my in-laws and especially my parents, as they make this journey through life possible and ever encourage me to achieve this goal.

Balancing Model Structure and Flexibility in Forecasting Financial Time Series

by

Jared Dale Fisher, Ph.D.
The University of Texas at Austin, 2019

Supervisor: Carlos M. Carvalho

This dissertation advances statistical methodology en route to providing new solutions to major questions in empirical finance. The common theme is the balance between structure and flexibility in these models. I show that structure, while it is potentially statistical bias, improves model performance when wisely chosen. Specifically, I look at asset returns' behavior: their relationship with firm characteristics, how they change over time, and what elements may cause their behavior.

First, I investigate the forecasting of multiple risk premia. Using the content of Fisher et al. (2019a), I introduce a simulation-free method to model and forecast multiple asset returns and employ it to investigate the optimal ensemble of features to include when jointly predicting monthly stock and bond excess returns. This approach builds on the Bayesian Dynamic Linear Models of West and Harrison (1997), and it can objectively determine, through a fully automated procedure, both the optimal set of regressors to include in the predictive system and the degree to which the model coefficients, volatilities, and covariances should vary over time. When applied to a portfolio of five stock and bond returns, I find that

my method leads to large forecast gains, both in statistical and economic terms. In particular, I find that relative to a standard no-predictability benchmark, the optimal combination of predictors, stochastic volatility, and time-varying covariances increases the annualized certainty equivalent returns of a leverage-constrained power utility investor by more than 500 basis points. Here, linear structure is chosen, and then I analyze what parameters should be flexible over time.

Second, I consider the problem of determining which characteristics of a firm impact its stock returns. Using the content of Fisher et al. (2019b), I first model a firm's expected return as a nonlinear, nonparametric function of its observable characteristics. I investigate whether theoretically-motivated monotonicity constraints on characteristics and nonstationarity of the conditional expectation function provide statistical and economic benefit. Then, using this model, I provide an approach for characteristic selection using utility functions to summarize the posterior distribution. Standard unexplained volume, short-term reversal, size, and variants of momentum are found to be significant characteristics, and there is evidence that this set changes in time. The data also provide strong support for monotonicity in some of the characteristics' relationships with returns. Hence, the flexibility of the nonlinear, nonparametric curves are regulated by monotonic constraints.

Finally, I turn to causal inference to ask which of these characteristics have causal relationships with asset returns. Hahn et al. (2018b) allow for regularized estimation of heterogeneous effects, and I modify their work to allow for non-binary, continuous treatments. This method is highly flexible at fitting complicated response surfaces with discontinuities, interactions, and nonlinearities, and thus benefits from added structure in the form of regularization from shrinkage priors. I demonstrate the model's ability to show the effect of firm size on returns, while controlling for book-to-market.

Table of Contents

List of Tables

List of Figures

Chapter 1

Optimal Asset Allocation with Multivariate Bayesian Dynamic Linear Models

This chapter is based on the text and content in Fisher et al. (2019a). We forecast future excess returns (risk premia) for multiple risky assets. Linear models are assumed as the structure, and we examine which components should be flexible over time, such as time-varying parameters and stochastic volatility.

1.1 Introduction

The study of portfolio theory and its implications for the asset allocation decisions of investors has and continues to play a central role in financial economics. Within this literature, a highly debated item over the years has been the question of whether asset returns are predictable and the extent to which this predictability affects the investor's optimal allocation choices.

There is by now an extensive empirical literature that has found evidence for predictability in stock and bond returns by means of valuation ratios, interest rates, and macroeconomic quantities.[1] Prior to the turn of the century, much of this literature focused on identifying variables that had significant and robust in-sample predictive power when forecasting returns. However, thanks in part to evidence uncovered in studies such as Bossaerts and Hillion (1999), Ang and Bekaert (2007), and Welch and Goyal (2008), in recent years

[1] See for example Fama and Schwert (1977) this paper. Campbell and Shiller (1988), Lettau and Ludvigson (2001), Lewellen (2004), and Ang and Bekaert (2007).

the emphasis has been gradually shifting from in-sample predictability of stock returns to out-of-sample predictability. A similar pattern has been observed for bond returns, where Thornton and Valente (2012) have shown that the information subsumed into forward rates and forward spreads, while quite successful in-sample, does not generate systematic economic value to investors out-of-sample.

The disparities between in-sample and out-of-sample evidence of return predictability can be in part explained by the presence of model instability in return prediction models. Due to the regular occurrence of a multitude of shocks to financial markets and the overall economy, investors are facing a constantly evolving, uncertain landscape and need to resort to highly adaptive methods when building their forecasts. By now, it is clear that not a single feature alone, but an ensemble of features is required to cope with the resulting uncertainty and instability as well as generate good predictions. This has been shown to be true for stock returns (Johannes et al., 2014) as well as for bond returns (Gargano et al., 2017). In particular, features that satisfy these out-of-sample needs include model and parameter uncertainty, time-varying volatility, time-varying parameters, and economically motivated constraints.

While there is ample evidence backing said ensemble of features when modeling returns on a single risky asset, surprisingly no study has examined how these features interact when jointly forecasting the returns of multiple risky assets. Yet, most investors hold many risky assets at once in their portfolios, which makes this an empirically relevant question. The primary contribution of this paper is to unify the features highlighted in the aforementioned papers into a single, computationally friendly framework capable of jointly handling multiple risky assets from different classes. Specifically, our framework builds on the Bayesian Dynamic Linear Models (DLMs) of West and Harrison (1997) and Gruber and West (2016) and examines a Bayesian agent who recursively updates her prior beliefs as new data is observed, therefore mimicking the real time decision making process of an investor.

The key element of our modeling approach is the ability to integrate a number of useful features into a flexible yet computationally simple method. First, our approach is well suited to integrate parameter uncertainty into the problem, as the DLMs yield predictive densities, rather than point forecasts, for each asset return. Second, the DLM framework allows for multivariate stochastic volatility. Both Johannes et al. (2014) and Gargano et al. (2017) find that stochastic volatility is a key feature to incorporate when modeling and forecasting stock and bond returns. The benefits of stochastic volatility are particularly pronounced during periods of very high market turmoil, such as the dot-com bubble as well as the most recent financial crisis. Given our emphasis on jointly modeling multiple risky assets, the key adjustment herein is how we model time variation in the cross-asset covariances. We provide two alternative approaches to handle this. Our first method builds on the Wishart DLM (W-DLM, henceforth) of West and Harrison (1997). Two key restrictions of the W-DLM are that, first, it forces all the assets in the system to share the same vector of predictor variables, and second, that variances and covariances are modeled in the same structure and must time-varying jointly. While in some settings this requirement may be appropriate, it is likely not a desirable feature when working with returns from very heterogeneous asset classes, such as equity and fixed income. To alleviate these concerns, our second approach builds on the Simultaneous Graphical DLM (SG-DLM, henceforth) of Gruber and West (2016). The SG-DLM permits each asset to feature its own set of predictor variables. In addition, the SG-DLM can easily be modified to allow for separate degrees of time variation for variances and covariances, which we find to be a very useful feature with financial returns. Most importantly, both DLM methods, as we present them, yield closed-form solutions for all the moments of the posterior distributions and predictive densities, and hence are computationally faster than the particle filter algorithm of Johannes et al. (2014) or the Markov chain Monte Carlo approaches of Gargano et al. (2017) and others.[2]

[2]To have its forward filter be closed-form, SG-DLMs must assume an appropriate dependence structure

Third, our models allow for time variation in the regression coefficients. It has been shown extensively that the regression coefficients of asset return predictive regressions change over time (Viceira, 1997; Pastor and Stambaugh, 2001; Kim et al., 2005; Paye and Timmermann, 2006; Lettau and Van Nieuwerburgh, 2008; Pettenuzzo and Timmermann, 2011). Rather than allowing for discrete non-recurring shifts, we let the regression coefficients evolve over time by adopting a flexible time-varying parameters specification. In this regard, our work is similar to Dangl and Halling (2012), who model and forecast the S&P 500 index and find that time-varying parameter models are strongly preferable to predictive regression with constant coefficients.

Fourth, our approach controls for model uncertainty through model averaging. Specifically, we combine forecasted densities from many models, as investigated by Rapach et al. (2010), Billio et al. (2013), and Pettenuzzo and Ravazzolo (2016). Thanks to the computational savings afforded by our approach, we are able to consider in reasonable computation time both the uncertainty regarding the degree to which parameters, volatilities, and covariances vary over time, as well as which predictors should be included in the model. We accomplish this by first fitting a separate DLM to each possible permutation of predictors and degrees of time variation. Next, we compute the predictive densities implied by each of these permutations and combine them together using both equal-weighted and score-weighted combinations. The latter weights the different model permutations according to their historical statistical fit, as measured by their logarithmic predictive scores.

Our secondary contribution is to empirically test the roles played by these features when forecasting multiple stock and bond returns. More specifically, we evaluate the performance of the W-DLM and SG-DLM models by jointly modeling the monthly excess returns on the five- and ten-year Treasury bonds, as well as the excess returns on the size-sorted small-, mid-, and large-cap stock portfolios. As for the predictors, we include the 15 variables

across the asset returns in the system. Details are given in Subsection 1.2.2.

studied in Welch and Goyal (2008) as well as the three predictors for bond returns considered by Gargano et al. (2017), namely forward spreads, the Cochrane and Piazzesi (2005) factor and the Ludvigson and Ng (2009) factor. We then estimate a W-DLM and a SG-DLM for each different combination of stock and bond predictors as well as combinations of different degrees of time variation in the regression coefficients, variances, and covariances. These individual DLMs are then averaged together in different groups to account for the aforementioned model uncertainty.

We evaluate the predictive performance of the various models and features over the 1985-2014 period against a simple no-predictability benchmark, and we find large statistical and economic benefits from using the appropriate ensemble of features. Among the features we consider, we find that W-DLMs and SG-DLMs with stochastic volatility bring the largest gains in terms of statistical predictability. In terms of economic predictability, which we quantify using certainty equivalent returns, we find that the optimal set of features includes SG-DLMs with stochastic volatility and time-varying covariances. In particular, we find that when using the optimal set of features our leverage-constrained power utility investor earns over 500 basis points (on an annualized basis) more than if she relied on the no-predictability benchmark.

Our paper relates to several branches of the literature. The papers most closely related to this paper are Dangl and Halling (2012), Johannes et al. (2014), and Gargano et al. (2017). All three papers focus on modeling and forecasting asset returns (stocks in the first two cases, treasury bonds in the last case) using flexible model specifications and building density forecasts that are robust to the presence of model instability and model uncertainty. In particular, Dangl and Halling (2012) use a DLM that is similar to what we employ here and allow for model uncertainty over different predictors and degrees of time variation in the regression coefficients (but do not allow for stochastic volatility). In contrast, Johannes et al. (2014) and Gargano et al. (2017) allow for both time-varying regression coefficients

and stochastic volatility, but because of their reliance on MCMC methods are forced to set *a priori* the degree to which parameters and volatility change over time. In addition, all three papers focus on univariate models and forecast a single financial asset at a time.[3] Relative to their setup, our approach jointly models multiple risky assets and takes into account the model uncertainty that arises from the availability of multiple predictors and from not knowing the degree of time variation in the regression coefficients, variances, and covariances.

There is also a small literature that has focused on forecasting multiple risky assets from different asset classes. Brennan et al. (1997) look at a portfolio that includes a stock index, a bond index, and cash, and forecast each asset return using a distinct predictor. This leads to a number of computational complexities, which they solve by estimating partial differential equations numerically. Wachter and Warusawitharana (2009) model the returns of both a stock index and a long-term bond using a single predictor variable but, because of their specific setup, need to rely on MCMC methods. Gao and Nardari (2018) model the returns of stocks, bonds, cash, and commodities by fitting multiple models with single predictors and averaging them with equal weights. They also allow for a time-varying covariance matrix, which they implement via the dynamic conditional correlation method of Engle (2002). Relative to these papers, ours provides the first attempt to objectively determine the optimal combination of features to include when modeling multiple risky assets at once, and does so by using a computationally efficient and simulation-free approach.

The remainder of the paper is organized as follows. Section 1.2 introduces the W-DLM and SG-DLM model specifications, the set of features we control for and our approach for averaging across all permutations of predictors and model characteristics. Next, Section

[3]While the main focus in Gargano et al. (2017) is on univariate predictive regressions, they include an application where they extend their setup to forecasting multiple treasury bond returns (differing in their maturities) at once.

1.3 describes the data and priors we adopted, while Section 1.4 summarizes our empirical analysis and the results we obtain. Finally, Section 1.5 provides some concluding remarks.

1.2 Our Approach

In this section, we introduce the approach we rely on to estimate and forecast multiple risky asset returns. We begin by describing in Subsections 1.2.1 and 1.2.2 the two Bayesian dynamic linear models (DLMs) we work with, namely the Wishart Dynamic Linear Model and the Simultaneous Graphical Dynamic Linear Model. Both methods allow the regression coefficients, variances, and covariances to vary over time and are therefore capable of coping with the model instability that plagues the relationship between asset returns and predictor variables. At the same time, both methods require the investor to know *a priori* the degree of time variation in the model parameters as well as the right combination of predictors to include in the regressions. In practice, the investor is likely unaware of what the optimal predictive model may look like, and is therefore facing uncertainty across all these dimensions. In Subsection 1.2.3, we describe a fully-automated data-based approach that we use to resolve this uncertainty.

1.2.1 Wishart Dynamic Linear Model

One of the key advantages of DLMs, compared to other Bayesian approaches, is that they feature closed-form solutions for all parameter updates as well as model forecasts. This is accomplished by a simulation-free procedure, known as a deterministic forward filter, which simulates how most people think, i.e. modifying their prior beliefs in real time as new data becomes available. More specifically, the posterior distribution of all model parameters at time $t-1$ becomes the prior at time t, and once time t data becomes available, a simple set of formulas merge time t priors and time t likelihood into time t posteriors. As part of this process, real time predictive densities and point forecasts can be obtained in a straightforward

manner. This procedure is repeated throughout the sample, thus yielding a sequence of posterior distributions and predictive densities.

Our first approach builds on the Wishart DLM (W-DLM) of West and Harrison (1997), which allows for time-varying regression coefficients as well as time-varying variances and covariances. As its name suggests, the W-DLM assumes that the error covariance matrix follows an inverse-Wishart distribution ($\mathcal{IW}$, henceforth). This is paired with the additional restriction that all the equations in the system share the same predictor variables.[4]

Let $\boldsymbol{r}_t$ denote a $q \times 1$ vector of log excess returns at time t $(t = 1, ..., T)$ and $\boldsymbol{x}_{t-1}$ represent a $p \times 1$ vector of lagged predictor variables, common to all q risky assets (throughout, we use bold lower-case letters to represent vectors and bold capitalized letters to represent matrices).[5] The W-DLM can be written as:

$$\boldsymbol{r}_t = \boldsymbol{B}_t' \boldsymbol{x}_{t-1} + \boldsymbol{v}_t \qquad\qquad \boldsymbol{v}_t | \boldsymbol{\Sigma}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_t) \tag{1.1}$$

where $\boldsymbol{B}_t$ is the $p \times q$ matrix of time-varying regression coefficients, which evolve over time according to pq random walk processes,

$$vec(\boldsymbol{B}_t) = vec(\boldsymbol{B}_{t-1}) + \boldsymbol{\omega}_t \qquad\qquad \boldsymbol{\omega}_t | \boldsymbol{\Sigma}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_t \otimes \boldsymbol{W}_t) \tag{1.2}$$

with $\boldsymbol{\omega}_t$ denoting a $pq \times 1$ vector of zero-mean normally distributed error terms, $vec\,(\cdot)$ is the vectorization operator, and $\otimes$ represents the Kronecker product.[6,7]

[4]The W-DLM is a generalization of the approach employed by Dangl and Halling (2012) to model and forecast stock returns. Relative to Dangl and Halling (2012), the W-DLM allows a modeler to model multiple risky assets at once and to include time-varying variances and covariances. It is essentially a deterministic forward-filter analog of the approach of Wachter and Warusawitharana (2009).

[5]$\boldsymbol{x}_{t-1}$ may or may not include a constant/intercept term.

[6]Specifically, the vectorization of an $m \times n$ matrix $\boldsymbol{A}$, denoted $vec(\boldsymbol{A})$, is the $mn \times 1$ column vector obtained by stacking the columns of the matrix $\boldsymbol{A}$ on top of one another.

[7]The W-DLM in Equation (1.1) can also be written using the matrix-normal distribution, i.e. $\boldsymbol{B}_t = \boldsymbol{B}_{t-1} + \boldsymbol{\Omega}_t, \quad \boldsymbol{\Omega}_t | \boldsymbol{\Sigma}_t \sim \mathcal{MN}(\boldsymbol{0}, \boldsymbol{W}_t, \boldsymbol{\Sigma}_t)$. Here $\boldsymbol{\Omega}_t$ follows a matrix-normal distribution $\mathcal{MN}$ with left variance matrix $\boldsymbol{W}_t$ and right variance matrix $\boldsymbol{\Sigma}_t$. This is the notation adopted by West and Harrison (1997, Section 16.2). See also Dawid (1981) for a description of the matrix-normal distribution and its properties.

Next, the $q \times 1$ error vector $\boldsymbol{v_t}$ is independently and normally distributed over time with variance-covariance matrix $\boldsymbol{\Sigma}_t$, given by

$$\boldsymbol{\Sigma}_t = \begin{bmatrix} \sigma_{1,t}^2 & \sigma_{12,t} & \dots & \sigma_{1q,t} \\ \sigma_{12,t} & \sigma_{2,t}^2 & \dots & \sigma_{2q,t} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1q,t} & \sigma_{2q,t} & \dots & \sigma_{q,t}^2 \end{bmatrix} \tag{1.3}$$

where both the variances $\left(\sigma_{1,t}^2, \dots, \sigma_{q,t}^2\right)$ and the covariances $\sigma_{ij,t}$ $(i, j = 1, \dots, q, j > i)$ are allowed to vary over time. Finally, the $p \times p$ matrix $\boldsymbol{W}_t$ controls the degree of time variation of the regression coefficient matrix $\boldsymbol{B}_t$, and we will specify its exact form below.

The model in Equations (1.1) - (1.3) is completed by specifying the initial states for both the regression coefficients and the variance-covariance matrix at time $t = 0$. These are given by the following distributions, where $\mathcal{D}_0$ denotes the information set available at time $t = 0$

$$\begin{aligned} vec(\boldsymbol{B}_0)|\boldsymbol{\Sigma}_0, \mathcal{D}_0 &\sim \mathcal{N}(vec(\boldsymbol{M}_0), \boldsymbol{\Sigma}_0 \otimes \boldsymbol{C}_0) \\ \boldsymbol{\Sigma}_0|\, \mathcal{D}_0 &\sim \mathcal{IW}(n_0, \boldsymbol{S}_0) \end{aligned} \tag{1.4}$$

Here, the $p \times q$ matrix $\boldsymbol{M}_0$ denotes the mean of the coefficient matrix $\boldsymbol{B}_0$, while the $p \times p$ matrix $\boldsymbol{C}_0$ summarizes the degree of confidence in $\boldsymbol{M}_0$. Similarly, $\boldsymbol{S}_0$ represents an estimate of the $q \times q$ error covariance matrix $\boldsymbol{\Sigma}_0$, which follows an Inverse-Wishart distribution with n_0 degrees of freedom. n_0, in turn, can be interpreted as the effective sample size of the initial state.

In practice, Equation (1.4) can also be interpreted as the posterior distribution of the parameters at time $t = 0$. We use this initial posterior in a process called *evolution*, where at any point in time t $(t = 1, ..., T)$ we use the posterior distribution from time $t - 1$ to compute the prior distribution of the parameters at time t. This is given by

$$\begin{aligned} vec(\boldsymbol{B}_t)|\boldsymbol{\Sigma}_t, \mathcal{D}_{t-1} &\sim \mathcal{N}(vec(\boldsymbol{M}_{t-1}), \boldsymbol{\Sigma}_t \otimes \hat{\boldsymbol{C}}_t) \\ \boldsymbol{\Sigma}_t|\, \mathcal{D}_{t-1} &\sim \mathcal{IW}(\hat{n}_t, \boldsymbol{S}_{t-1}) \end{aligned} \tag{1.5}$$

where $\hat{\boldsymbol{C}}_t$ and $\hat{n}_t$ are modified versions of $\boldsymbol{C}_{t-1}$ and n_{t-1} and are used as estimates of $\boldsymbol{C}_t$ and n_t. In particular, we set

$$\hat{\boldsymbol{C}}_t = \frac{1}{\delta_\beta}\boldsymbol{C}_{t-1} \tag{1.6}$$

and

$$\hat{n}_t = \delta_v n_{t-1} \tag{1.7}$$

where $\delta_\beta \in (0, 1]$ and $\delta_v \in (0, 1]$ denote discount factors. δ_β is incorporated into the model (and hence we can control the degree of time variation of the regression coefficient matrix $\boldsymbol{B}_t$) by rewriting the $p \times p$ matrix $\boldsymbol{W}_t$ in Equation (1.2) as

$$\boldsymbol{W}_t = \frac{1 - \delta_\beta}{\delta_\beta}\boldsymbol{C}_{t-1}, \tag{1.8}$$

which suggests that the smaller the discount factor δ_β is, the larger the elements of the covariance matrix $\boldsymbol{W}_t$ will be, thus increasing the variance/uncertainty around time t regression coefficients and allowing $\boldsymbol{B}_t$ to move further away from $\boldsymbol{B}_{t-1}$. In the extreme case of $\delta_\beta = 1$ we have that $\hat{\boldsymbol{C}}_t = \boldsymbol{C}_{t-1}$ and $\boldsymbol{W}_t = 0$, which means that when $\delta_\beta = 1$ the regression coefficient matrix $\boldsymbol{B}_t$ does not vary over time. As for δ_v, note that Equation (1.5) implies that

$$\mathbb{E}(\boldsymbol{\Sigma}_t | \mathcal{D}_{t-1}) = \frac{1}{\hat{n}_t - q - 1}\boldsymbol{S}_{t-1} \tag{1.9}$$

which means that the smaller δ_v is, the larger the expected value of all elements in the error covariance matrix will be. Also, it can be shown that for large t, $0 < \delta_v < 1$ implies that the posterior estimates of the variances and covariances across series essentially become exponentially weighted moving averages of the past sample variances and sample covariances, with weights that decay over time as a function of δ_v. This, in turn, suggests that the smaller the discount factor δ_v is, the quicker $\boldsymbol{\Sigma}_t$ can adapt to the new data and the more it can move away from $\boldsymbol{\Sigma}_{t-1}$. Finally, in the extreme case of $\delta_v = 1$, we obtain a model where there is

no discounting of the old data and thus $\boldsymbol{\Sigma}_t$ is assumed constant, i.e. a constant volatility model.

With Equation (1.5) in hand, it becomes possible to compute the predictive distribution of $\boldsymbol{r}_t$, conditional on the information set available at time $t-1$. In particular, we have that

$$\boldsymbol{r}_t|\delta_\beta, \delta_v, \mathcal{D}_{t-1} \sim \mathcal{T}_{\hat{n}_t}\left(\boldsymbol{M}'_{t-1}\boldsymbol{x}_{t-1}, \quad \boldsymbol{S}_{t-1}(1+\boldsymbol{x}'_{t-1}\hat{\boldsymbol{C}}_t\boldsymbol{x}_{t-1})\right). \tag{1.10}$$

where $\mathcal{T}_{\hat{n}_t}$ denotes a Student's t-distribution with $\hat{n}_t$ degrees of freedom.[8] This implies that the conditional forecast of the mean vector and variance-covariance matrix of $\boldsymbol{r}_t$ will be given by

$$\mathbb{E}[\boldsymbol{r}_t|\delta_\beta, \delta_v, \mathcal{D}_{t-1}] = \boldsymbol{M}'_{t-1}\boldsymbol{x}_{t-1} \tag{1.11}$$

$$Cov[\boldsymbol{r}_t|\delta_\beta, \delta_v, \mathcal{D}_{t-1}] = \frac{\hat{n}_t}{\hat{n}_t - 2}\boldsymbol{S}_{t-1}(1+\boldsymbol{x}'_{t-1}\hat{\boldsymbol{C}}_t\boldsymbol{x}_{t-1}). \tag{1.12}$$

After observing the actual returns for time period t, we can update the prior for time t from Equation (1.5) into the posterior for time t. We provide the details of the closed-form updating equations in Appendix A, where we show how from the initial states in Equation (1.4), the sequence of regression coefficients $\{\boldsymbol{B}_t\}_{t=1}^T$ and variance-covariance matrices $\{\boldsymbol{\Sigma}_t\}_{t=1}^T$ can be obtained by a simple and very fast forward filter. Thus, the W-DLM deterministically gives the posterior distribution of the model parameters at each time step, avoiding the need for computationally expensive Markov chain Monte Carlo simulation methods.

While computationally very fast, the W-DLM presents three key drawbacks. First, very much like Wachter and Warusawitharana (2009)'s model, the W-DLM uses the same predictors for each asset. The severity of this restriction will depend on the particular

[8]Note that we have opted for a notation where we make explicit the dependence of the predictive distribution for $\boldsymbol{r}_t$ (and its moments) to the choices made with respect to the two discount factors, δ_β and δ_v.

assets being modeled, but it is not hard to imagine situations where this restriction may not be desirable. Second, the conjugate inverse Wishart prior, while computationally very convenient, is notoriously inflexible and may not adapt well to underlying data.[9] Finally, by construction the W-DLM features a single discount factor for the entire covariance matrix, which means that both the variances and covariances will be discounted in the same way. In the next section, we present a more general approach that will permit us to relax all three drawbacks of the W-DLM.

1.2.2 Simultaneous Graphical Dynamic Linear Model

Our second approach builds on the simultaneous graphical dynamic linear model (SG-DLM) of Gruber and West (2016). Relative to the W-DLM method described in the previous section, one of the key advantages of the SG-DLM is that it can accommodate asset-specific regressors, while still allowing for time-varying regression coefficients, variances, and covariances. This is accomplished through a modeling strategy that "decouples" the joint dynamic system into separate univariate models for each of the risky assets, taking into full account the contemporaneous dependencies across assets. In turn, these univariate models can be updated with great computational speed, thus preserving the closed-form forward filter nature of the algorithm. We begin by re-writing the joint dynamic system for the q excess returns $\boldsymbol{r}_t$ as follows:

$$\boldsymbol{r}_t = \begin{pmatrix} \boldsymbol{x}'_{1,t-1}\boldsymbol{\beta}_{1t} \\ \vdots \\ \boldsymbol{x}'_{q,t-1}\boldsymbol{\beta}_{qt} \end{pmatrix} + \begin{pmatrix} \boldsymbol{r}'_{-1,t}\boldsymbol{\gamma}_{1t} \\ \vdots \\ \boldsymbol{r}'_{-q,t}\boldsymbol{\gamma}_{qt} \end{pmatrix} + \boldsymbol{\nu}_t \quad \boldsymbol{\nu}_t | \, \boldsymbol{\Omega}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_t) \tag{1.13}$$

where $\boldsymbol{x}_{j,t-1}$ $(j = 1, ..., q)$ denotes the $p_j \times 1$ vector of asset j's specific lagged predictors (possibly including an intercept), while $\boldsymbol{r}_{-j,t}$ represents the contemporaneous log excess returns of all assets other than asset j. Similarly, $\boldsymbol{\beta}_{jt}$ denotes the $p_j \times 1$ vector of the predictors'

[9] See for example Barnard et al. (2000) or Gelman and Hill (2006). One simple point is that the inverse Wishart has only a single parameter governing the variability about all of its elements, thus, for a distribution of a covariance matrix, your uncertainty about all the variances and covariances must be the same.

coefficients while $\boldsymbol{\gamma}_{jt}$ is the $(q-1)\times 1$ vector of coefficients capturing the contemporaneous correlations between asset j's log excess return and the remaining $q-1$ log excess returns. Finally, $\boldsymbol{\Omega}_t = diag\left(\sigma_{1t}^2, ..., \sigma_{qt}^2\right)$ is a $q \times q$ matrix with the assets' error variances on the diagonal.

Note that relative to the W-DLM specification in Equation (1.1), which models the contemporaneous correlations across asset returns through the full variance-covariance matrix $\boldsymbol{\Sigma}_t$, the system in Equation (1.13) handles the contemporaneous correlations by introducing the $\boldsymbol{\gamma}_{jt}$ parameters and the $\boldsymbol{r}_{-j,t}$ regressors $(j = 1, ..., q)$ while leaving all elements of the error term $\boldsymbol{\nu}_t$ contemporaneously uncorrelated, i.e. $\nu_{it} \perp \nu_{jt}$ for all $i \neq j$. This modeling choice, as we will show shortly, is what allows the SG-DLM to continue working with a closed-form forward-filter even after relaxing the restrictions enforced by the W-DLM.

We proceed by combining all elements of $\boldsymbol{\gamma}_{1t}$ to $\boldsymbol{\gamma}_{qt}$ into the $q \times q$ zero-diagonal matrix $\boldsymbol{\Gamma}_t$ as follows,

$$\boldsymbol{\Gamma}_t = \begin{bmatrix} 0 & \gamma_{12,t} & \cdots & \gamma_{1q-1,t} & \gamma_{1q,t} \\ \gamma_{21,t} & 0 & \cdots & \gamma_{2q-1,t} & \gamma_{2q,t} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_{q1,t} & \gamma_{q2,t} & \cdots & \gamma_{qq-1,t} & 0 \end{bmatrix} \tag{1.14}$$

which in turn allows us to rewrite Equation (1.13) as

$$\boldsymbol{r}_t = \begin{pmatrix} \boldsymbol{x}'_{1,t-1}\boldsymbol{\beta}_{1t} \\ \vdots \\ \boldsymbol{x}'_{q,t-1}\boldsymbol{\beta}_{qt} \end{pmatrix} + \boldsymbol{\Gamma}_t \boldsymbol{r}_t + \boldsymbol{\nu}_t \quad \boldsymbol{\nu}_t | \, \boldsymbol{\Omega}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Omega}_t). \tag{1.15}$$

It is easy to show that we can further rearrange Equation (1.15) to write

$$\boldsymbol{r}_t = (I - \boldsymbol{\Gamma}_t)^{-1} \begin{pmatrix} \boldsymbol{x}'_{1,t-1}\boldsymbol{\beta}_{1t} \\ \vdots \\ \boldsymbol{x}'_{q,t-1}\boldsymbol{\beta}_{qt} \end{pmatrix} + \boldsymbol{u}_t \quad \boldsymbol{u}_t | \, \boldsymbol{\Sigma}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_t) \tag{1.16}$$

where $\boldsymbol{\Sigma}_t = (I - \boldsymbol{\Gamma}_t)^{-1}\boldsymbol{\Omega}_t((I - \boldsymbol{\Gamma}_t)^{-1})'$ is now a full variance-covariance matrix capturing the contemporaneous correlations among the q assets. As shown by Gruber and West (2016), the

presence of the $(I-\mathbf{\Gamma}_t)^{-1}$ term in Equation (1.16) significantly complicates the inference, as the joint posterior of the parameters is now proportional to the determinant $|I-\mathbf{\Gamma}_t|$ times the product of q univariate normal densities, i.e.

$$p(\boldsymbol{r}_t|\boldsymbol{\beta}_{1t},...,\boldsymbol{\beta}_{qt},\mathbf{\Gamma}_t,\mathbf{\Omega}_t) \propto |I-\mathbf{\Gamma}_t| \prod_{j=1}^{q} p(r_{jt}|\boldsymbol{\beta}_{jt},\boldsymbol{\gamma}_{jt},\sigma^2_{jt}). \tag{1.17}$$

The obvious exception to this rule is the case where $|I-\mathbf{\Gamma}_t|=1$. In this case, as we will show below, it becomes possible to derive the multivariate distribution of all assets using fast and reliable univariate forward filters similar to those introduced by West and Harrison (1997). This is indeed the avenue we explore here.[10]

In particular, we follow Primiceri (2005), Carriero et al. (2016), and Koop et al. (2018) and assume that the dynamic system in Equation (1.13) is fully recursive. This, in turn, implies that the $\mathbf{\Gamma}_t$ matrix in Equation (1.15) becomes lower triangular, still featuring zeros on its main diagonal. Next, we write r_{jt}, the log excess return of risky asset j at time t, as a linear combination of a $p_j \times 1$ vector of asset-specific lagged predictors $\boldsymbol{x}_{j,t-1}$ as well as the contemporaneous log excess returns from the previous $j-1$ assets, which we denote with $\boldsymbol{r}_{<j,t}$,

$$r_{jt} = \boldsymbol{x}'_{j,t-1}\boldsymbol{\beta}_{jt} + \boldsymbol{r}'_{<j,t}\boldsymbol{\gamma}_{<j,t} + \nu_{jt} \qquad\qquad \nu_{jt} \sim N(0,\sigma^2_{jt}) \tag{1.18}$$

where $\boldsymbol{\gamma}_{<j,t}$ is the $(j-1)\times 1$ vector of coefficients associated with the contemporaneous excess returns $\boldsymbol{r}_{<j,t}$. We now specify the law of motion for the regression coefficients $\boldsymbol{\beta}_{jt}$ and $\boldsymbol{\gamma}_{<j,t}$:

$$\begin{pmatrix}\boldsymbol{\beta}_{jt}\\ \boldsymbol{\gamma}_{<j,t}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\beta}_{j,t-1}\\ \boldsymbol{\gamma}_{<j,t-1}\end{pmatrix} + \boldsymbol{\omega}_{jt} \qquad\qquad \boldsymbol{\omega}_{jt} \sim N(\mathbf{0},\boldsymbol{W}_{jt}). \tag{1.19}$$

where $\boldsymbol{\omega}_{jt}$ is the $(p_j+j-1)\times 1$ vector of evolution errors with covariance matrix $\boldsymbol{W}_{jt}$.

[10] In particular, we build on Zhao et al. (2016), who present a forward filter algorithm for a dynamic linear system with a fully recursive triangular specification (where Equation (1.14) is a triangular matrix), where the parameters within each equation of the system are updated individually.

The SG-DLM is completed by specifying the initial states of the model parameters, that is, regression coefficients $\boldsymbol{\beta}_{jt}$, contemporaneous returns coefficients $\boldsymbol{\gamma}_{<j,t}$, and variance term σ^2_{jt}. For each asset j, we write

$$\begin{aligned}\begin{pmatrix}\boldsymbol{\beta}_{j0}\\ \boldsymbol{\gamma}_{<j,0}\end{pmatrix}\Bigg|\, \sigma^2_{j0}, \mathcal{D}_0 &\sim \mathcal{N}\left(\boldsymbol{m}_{j0},\ \frac{\sigma^2_{j0}}{s_{j0}}\boldsymbol{C}_{j0}\right)\\ \sigma^{-2}_{j0}|\mathcal{D}_0 &\sim \mathcal{G}\left(\frac{n_{j0}}{2}, \frac{n_{j0}s_{j0}}{2}\right)\end{aligned} \tag{1.20}$$

where $\boldsymbol{m}_{j0}$ is a $(p_j+j-1)\times 1$ vector denoting the mean of the coefficients $\left(\boldsymbol{\beta}'_{j0}, \boldsymbol{\gamma}'_{<j,0}\right)'$, while $\boldsymbol{C}_{j0}$ is a $(p_j + j - 1) \times (p_j + j - 1)$ covariance matrix factor summarizing the uncertainty surrounding the mean estimates $\boldsymbol{m}_{j0}$. The initial error precision $1/\sigma^2_{j0}$ follows a Gamma distribution with mean $1/s_{j0}$ and degrees of freedom n_{j0}. n_{j0} can be interpreted as the effective sample size of this initial posterior. We further abbreviate these two distributions using the joint Normal-Gamma distribution

$$\begin{pmatrix}\boldsymbol{\beta}_{j0}\\ \boldsymbol{\gamma}_{<j,0}\end{pmatrix}, \sigma^2_{j0}\Bigg|\, \mathcal{D}_0 \sim \mathcal{NG}(\boldsymbol{m}_{j0},\ \boldsymbol{C}_{j0},\ n_{j0},\ s_{j0}). \tag{1.21}$$

As with the initial conditions for the W-DLM in Equation (1.4), Equation (1.21) can be interpreted as the posterior distribution of the model parameters at time $t = 0$. Once this process is initialized, at any given point in time t ($t = 1, ..., T$) we can use the posterior distribution from time $t - 1$ to compute the prior distributions of the model parameters at time t. These are given by

$$\begin{pmatrix}\boldsymbol{\beta}_{jt}\\ \boldsymbol{\gamma}_{<j,t}\end{pmatrix}, \sigma^2_{jt}\Bigg|\, \mathcal{D}_{t-1} \sim \mathcal{NG}(\boldsymbol{m}_{j,t-1}, \hat{\boldsymbol{C}}_{jt}, \hat{n}_{jt}, s_{j,t-1}). \tag{1.22}$$

where $\hat{\boldsymbol{C}}_{jt}$ and $\hat{n}_{jt}$ are modified versions of $\boldsymbol{C}_{j,t-1}$ and $n_{j,t-1}$, and are used as estimates of $\boldsymbol{C}_{j,t}$ and $n_{j,t}$. In particular, we set

$$\hat{\boldsymbol{C}}_{j,t} = \begin{bmatrix}\boldsymbol{C}_{\beta\beta j,t-1}/\delta_{\beta j} & \boldsymbol{C}_{\beta\gamma j,t-1}\\ \boldsymbol{C}_{\gamma\beta j,t-1} & \boldsymbol{C}_{\gamma\gamma j,t-1}/\delta_{\gamma j}\end{bmatrix}. \tag{1.23}$$

and

$$\hat{n}_{jt} = \delta_{vj} n_{j,t-1} \tag{1.24}$$

where $\delta_{\beta j} \in (0,1]$, $\delta_{\gamma j} \in (0,1]$, and $\delta_{vj} \in (0,1]$ denote asset-specific discount factors. In particular, as shown in Equation (1.23), the updated variance term $\hat{\boldsymbol{C}}_{j,t}$ features different blocks, separating asset j's predictor coefficients $\boldsymbol{\beta}_{jt}$ from asset j's correlation factors $\boldsymbol{\gamma}_{<j,t}$. In turn, this gives the user the freedom to introduce, asset by asset, a separate discount factor for the correlations ($\delta_{\gamma j}$) and the predictor coefficients ($\delta_{\beta j}$), allowing each asset's dynamic regression coefficients and correlation factors to evolve over time at potentially different paces.[11] It is possible to show that

$$\boldsymbol{W}_{jt} = \begin{bmatrix} (\frac{1}{\delta_{\beta j}} - 1)\boldsymbol{C}_{\beta\beta j,t-1} & 0 \\ 0 & (\frac{1}{\delta_{\gamma j}} - 1)\boldsymbol{C}_{\gamma\gamma j,t-1} \end{bmatrix} \tag{1.25}$$

which suggests that the smaller the discount factors $\delta_{\beta j}$ and $\delta_{\gamma j}$ are, the larger the elements in the respective blocks of the covariance matrix $\boldsymbol{W}_{jt}$ will be, thus increasing the chances that $\boldsymbol{\beta}_{jt}$ and $\boldsymbol{\gamma}_{<j,t}$ will move further away from $\boldsymbol{\beta}_{j,t-1}$ and $\boldsymbol{\gamma}_{<j,t-1}$.[12] As for δ_{vj}, much like δ_v with the W-DLM, we have that small values of δ_{vj} lead to large variability (and thus flexibility) in the volatilities, with σ^2_{jt} allowed to to move further away from $\sigma^2_{j,t-1}$. In contrast, when $\delta_{vj} = 1$ there is no discounting of past data and, as a result, σ^2_{jt} does not vary over time.[13]

Once Equation (1.22) is available, it becomes possible to derive the predictive distribution for $\boldsymbol{r}_t$, conditional on the information set available at time $t-1$. Thanks to the fully recursive identification strategy we adopted, we can proceed sequentially through the

[11] This mimics the block discounting approach introduced by (West and Harrison, 1997, Section 6.3.2).

[12] Note that the zero off-diagonal blocks in Equation (1.25) represent an assumption (stemming from West and Harrison (1997, Section 6.3.2)), namely that the correlations between the predictor coefficients $\boldsymbol{\beta}_{jt}$ and the correlation factors $\boldsymbol{\gamma}_{<j,t}$ are constant (but not zero). This assumption, in turn, leads to having no discount factors in the denominators of the off-diagonal blocks of $\hat{\boldsymbol{C}}_{j,t}$ in Equation (1.23).

[13] We note that while in principle the SG-DLM permits each asset to have its own degree of time variation in coefficients, variances, and covariances, it is also quite easy to introduce restrictions in the model setup. For example, one could imagine a situation where all assets within a given class (e.g., bonds or stocks) share the same discount factors, or even a situation where, as it was the case with the W-DLM, all the assets in the system share the same discount factors.

q equations of the dynamic system. Starting with the first asset in the system, we have that

$$\mathbb{E}[r_{1t}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}] = \boldsymbol{x}'_{1,t-1}\boldsymbol{m}_{1,t-1} \tag{1.26}$$

$$Var[r_{1t}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}] = \frac{\hat{n}_{1t}}{\hat{n}_{1t}-2}(\boldsymbol{x}'_{1,t-1}\hat{\boldsymbol{C}}_{1t}\boldsymbol{x}_{1,t-1} + s_{1,t-1}). \tag{1.27}$$

where, as with the W-DLM forecasts, we have highlighted the dependence of these predictive moments on the choices made regarding the discount factors, that is $\boldsymbol{\delta}_j = (\delta_{\beta j}, \delta_{\gamma j}, \delta_{vj})$. As for the generic asset j in the system ($1 < j \leq q$), we begin by separating the elements of the coefficient mean vector $\boldsymbol{m}_{j,t-1}$ according to whether they relate to the lagged predictor variables or the contemporaneous returns, i.e. $\boldsymbol{m}_{j,t-1} = \left(\boldsymbol{m}'_{\beta_{j,t-1}}, \boldsymbol{m}'_{\gamma_{<j,t-1}}\right)'$.[14] It then follows that

$$\mathbb{E}[r_{jt}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}] = \boldsymbol{x}'_{j,t-1}\boldsymbol{m}_{\beta_{j,t-1}} + \mathbb{E}[\boldsymbol{r}_{<j,t}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}]'\boldsymbol{m}_{\gamma_{<j,t-1}}, \tag{1.28}$$

$$\begin{aligned} Var[r_{jt}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}] = \frac{\hat{n}_{jt}}{\hat{n}_{jt}-2}\left\{tr\left(\hat{\boldsymbol{C}}_{\gamma_{<j,t}} Cov[\boldsymbol{r}_{<j,t}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}]\right) + c_{jt} + s_{j,t-1}\right\} \\ + \boldsymbol{m}'_{\gamma_{<j,t-1}} Cov[\boldsymbol{r}_{<j,t}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}]\boldsymbol{m}_{\gamma_{<j,t-1}} \end{aligned} \tag{1.29}$$

and

$$Cov[r_{jt}, \boldsymbol{r}_{<j,t}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}] = \boldsymbol{m}'_{\boldsymbol{\gamma}_{<j,t-1}} Cov[\boldsymbol{r}_{<j,t}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}] \tag{1.30}$$

where $\mathbb{E}[r_{<jt}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}]$ and $Cov[\boldsymbol{r}_{<j,t}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}]$ are known, $tr()$ stands for the trace of a matrix, and

$$c_{jt} = \begin{pmatrix} \boldsymbol{x}_{j,t-1} \\ \mathbb{E}[\boldsymbol{r}_{<j,t}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}] \end{pmatrix}' \hat{\boldsymbol{C}}_{jt} \begin{pmatrix} \boldsymbol{x}_{j,t-1} \\ \mathbb{E}[\boldsymbol{r}_{<j,t}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}] \end{pmatrix} \tag{1.31}$$

Applied iteratively, Equations (1.27) - (1.30) yield the mean vector and covariance matrix of the predictive density of $\boldsymbol{r}_t$.

[14] In particular, $\boldsymbol{m}_{\beta_{j,t-1}}$ denotes the $p_j \times 1$ vector of coefficients for predictor variables $\boldsymbol{x}_{j,t-1}$, while $\boldsymbol{m}_{\gamma_{<j,t-1}}$ is the $(j-1) \times 1$ vector of coefficients for the vector of contemporaneous returns $\boldsymbol{r}_{<j,t}$.

After observing the actual returns for time period t, we can update the prior for time t into the posterior for time t. We provide the details of these updating formulas in Appendix B, where we show how, from the initial states of Equation (1.20), the sequence of regression coefficients and variance-covariance matrices can be obtained by a simple and very fast forward filter. Thus, the SG-DLM, like the W-DLM, deterministically gives the posterior distribution of the model parameters and the predictive densities of the q risky assets at each time step and avoids the need for computationally expensive Markov chain Monte Carlo methods.

1.2.3 Model Averaging

As we mentioned at the outset, both the W-DLM and the SG-DLM require the investor to know *a priori* the degree of time variation in the model parameters as well as the right combination of predictors to include in the model. In practice, the investor is unaware of what the optimal combination of these features may look like, and she is therefore facing significant uncertainty along these dimensions. To address this issue, we turn to model combinations.

For both the W-DLM and SG-DLM specifications, we estimate a different version of each model for every possible combination of predictor variables and discount factors. We defer the discussion of the predictors to the next section, where we will provide a detailed list of all the stock and bond predictors we consider in this study. As for the discount factors for the W-DLM, we consider values from two equally-spaced grids: $\delta_\beta, \in \{0.98, 0.99, 1.0\}$ and $\delta_v \in \{0.95, 0.975, 1.0\}$.[15] For the SG-DLM, we consider values from three equally-spaced grids, namely δ_β, $\delta_\gamma \in \{0.98, 0.99, 1.0\}$, and $\delta_v \in \{0.95, 0.975, 1.0\}$. We have dropped the

[15]While we could explore more of the model space by increasing the number of points used within these ranges, three values of each suffice to demonstrate the effects of model averaging and time variation. We find no notable changes when increasing to ten values within each grid. Likewise, Dangl and Halling (2012) use $\delta_\beta \in \{0.96, 0.98, 1.00\}$, and find no notable changes by doubling the granularity to $\delta_\beta \in \{0.96, 0.97, 0.98, 0.99, 1.00\}$.

j subscript here to indicate that, in our empirical application, all assets in a particular SG-DLM will share the same discount factors.[16]

Next, at each point in time, we combine the forecast distributions obtained from all the permutations of predictors and discount factors. We do this separately for both the W-DLM and SG-DLM models. Note that, while we could have also chosen to combine the resulting predictive densities across the two model specifications, we have elected to keep the two methods separated to better isolate the impact of the aforementioned W-DLM restrictions, and to empirically quantify the importance of relaxing such constraints with the SG-DLM approach. Also, in an attempt to slightly ease the notation, below we will use $\mathcal{M}_i$ to denote the model with the i-th permutation of predictors and discount factors considered, where $i = 1, ..., K_W$ in the case of the W-DLMs and $i = 1, ..., K_{SG}$ in the case of the SG-DLMs, and K_W (K_{SG}) denotes the total number of model permutations we consider. We will then generally refer to the time t predictive mean and covariance matrix that come out of the i-th permutation of predictors and discount factors with $\mathbb{E}(\boldsymbol{r}_t|\mathcal{M}_i, \mathcal{D}_{t-1})$ and $Cov(\boldsymbol{r}_t|\mathcal{M}_i, \mathcal{D}_{t-1})$.

We explore two alternative combination schemes, as both have seen empirical success in the stock and bond predictability literatures. Our first combination scheme allows the weights on individual forecasting models to reflect their past predictive accuracy, and is therefore inspired by the optimal prediction pool approach of Geweke and Amisano (2011) and its good performance in settings similar to ours, as documented by Pettenuzzo et al. (2014) and Gargano et al. (2017). Specifically, at each point in time t, we compute model $\mathcal{M}_i$'s weight ($i = 1, ..., K_W$ in the case of the W-DLMs and $i = 1, ..., K_{SG}$ in the case of

[16] As we mentioned before, the SG-DLM can allow each asset to have its own set of discount factors. However, for the specific empirical application considered in this paper we have found that a model with separate discount factors for each asset class does not outperform the simpler specification where the discount factors are constant across assets. Therefore, in what follows we will restrict our attention to the special case where $\delta_{\beta 1} = ... = \delta_{\beta q}$, $\delta_{\gamma 1} = ... = \delta_{\gamma q}$, and $\delta_{v1} = ... = \delta_{vq}$.

the SG-DLMs) by looking at its historical statistical performance up through time $t-1$, as determined by the multivariate log score:

$$w_{i,t} \propto \sum_{\tau=1}^{t-1} \ln(S_{i,\tau}) \tag{1.32}$$

Here $S_{i,\tau}$ denotes the recursively computed score for model i at time τ, which we obtain by evaluating a Gaussian density with mean vector and covariance matrix equal to $\mathbb{E}(\boldsymbol{r}_\tau|\mathcal{M}_i, \mathcal{D}_{\tau-1})$ and $Cov(\boldsymbol{r}_\tau|\mathcal{M}_i, \mathcal{D}_{\tau-1})$ at the realized log excess returns $\boldsymbol{r}_\tau$. This approach rewards the high-performing combinations of predictors and discount factors, assigning them more weight in the model combination. Our second combination scheme is the equal-weighted pool, which weight each of the K_W (or K_{SG}) models equally and has been shown by Rapach et al. (2010) to work well at least in the case of stock returns.

1.3 Data and Priors

1.3.1 Data

This section describes how we construct our portfolio of risky assets as well as which predictors we consider in our analysis.

1.3.1.1 Asset Returns

As for our pool of risky assets, we focus on a portfolio of monthly stock and bond returns, and, in particular, we consider: (i) the value-weighted return of the largest 20% of firms listed on the Center for Research in Security Prices's database (CRSP); (ii) the value-weighted return of the CRSP firms in between the median and 80th percentile in size; (iii) the value-weighted return of the smallest 50% of CRSP firms; (iv) the five-year Treasury bond return; (v) the 10-year Treasury bond return. In addition, we collect data on the one-month Treasury bill rate (from Ibbotson Associates), which we use in our analysis to denote the returns of a risk-free investment strategy and to compute excess returns. All returns

are continuously compounded, and the stock returns come from the CRSP's monthly cap-based portfolios file.[17] In contrast, monthly returns on five- and ten-year Treasury bonds are computed using the two-step procedure described in Gargano et al. (2017). In particular, in the first step we start from the daily yield curve parameter estimates of Gurkaynak et al. (2007) and use them to reconstruct the entire yield curve at the daily frequency. Next, focusing on the last day of each month's estimated log yields, we combine the interpolated log yields to generate non-overlapping monthly bond returns for various maturities.[18] Excess returns are obtained by subtracting the continuously compounded monthly T-bill rate from the previously computed asset returns.

1.3.1.2 Predictors

As for the predictors considered in this analysis, we start by including the equity predictors studied in Welch and Goyal (2008).[19] These variables can be divided into three groups, namely stock, treasury, and corporate bond market variables. Stock market variables include the dividend-price ratio, dividend-payout ratio, stock variance, book-to-market ratio, and net equity expansion. Treasury market variables include the Treasury bill rate, long-term yield, term spread, and inflation rate. Finally, the default yield spread incorporates information from the corporate bond market. We augment this list of variables with the three predictors for bond returns considered by Gargano et al. (2017). Specifically, we consider forward spreads as proposed by Fama and Bliss (1987), a linear combination of forward rates

[17]The T-bill rate comes from the research factors file, which is made available by Kenneth French at http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/datalibrary.html.

[18]Let n be the bond maturity in years. Time $t+1$ holding period bond returns are given by the following formula

$$r_{t+1}^{(n)} = n y_t^{(n)} - (n - h/12) y_{t+1}^{(n-h/12)},$$

where $y_t^{(n)}$ is the log yield of the time t bond with n periods to maturity. To obtain five- and ten-year bond returns, we set $n = 60$ and $n = 120$ respectively.

[19]We refer to Welch and Goyal (2008) for a detailed description of the construction of the individual predictors, which are available at http://www.hec.unil.ch/agoyal/.

as proposed by Cochrane and Piazzesi (2005), and a linear combination of macro factors, as proposed by Ludvigson and Ng (2009). We give both an equity and bond predictor to each DLM, which, for 15 stock predictors and three bond predictors, yields 45 different combinations. Hence, we specify 45 DLMs per combination of discount factors, yielding a total of 405 W-DLMs (due to $3^2 = 9$ different combinations of discount factors per each of the 45 predictor combinations) and 1215 SG-DLMs (from $3^3 = 27$ combinations of discount factors).[20]

Once combined, our sample of monthly excess returns and predictors spans from January 1962 to December 2014, for a total of 636 observations (635 observations once we lag the predictor values). We provide summary statistics for both excess returns and predictors in Table 1.1.

1.3.2 Initial States

As we described in Section 1.2, we use $\mathcal{D}_0$ to denote the information set that we rely on to initialize the W-DLM and SG-DLM forward filters. We set aside the first 120 months of data to initialize/train our models, hence $\mathcal{D}_0$ in our case denotes the time period ranging from January 1962 to January 1972. We center the initial states for both the W- and SG-DLM specifications on the models' OLS estimates obtained over $\mathcal{D}_0$. Specifically, in the W-DLM we set $\boldsymbol{M}_0$, the conditional mean of the initial state in Equation (1.4), to the coefficient estimates from an OLS multivariate predictive regression over the training dataset, and set $\boldsymbol{S}_0$ to the corresponding sample covariance matrix of the OLS residuals. Next, we specify $\boldsymbol{C}_0 = 100 I_p$, which effectively renders the prior on the initial state $\boldsymbol{B}_0$ uninformative. Finally, we set the degrees of freedom n_0 to 10, therefore down-weighting the prior on $\boldsymbol{\Sigma}_0$ and rendering it flat and uninformative.

[20]To make the comparison across models easier, each equation in a predictive system/DLM will include the same two predictors, that is, one bond and one stock predictor, together for all assets. This will be true regardless of whether we work with the W- or SG-DLM methods.

Table 1.1: Summary Statistics

	Mean	StDev	Min	P_{25}	P_{75}	Max	SR
Panel A: Excess Returns							
5 Year Bond	0.002	0.018	-0.089	-0.007	0.011	0.094	0.340
10 Year Bond	0.002	0.032	-0.119	-0.014	0.019	0.163	0.240
Large-Cap Stocks	0.004	0.043	-0.235	-0.019	0.031	0.153	0.293
Mid-Cap Stocks	0.005	0.052	-0.294	-0.023	0.040	0.200	0.348
Small-Cap Stocks	0.006	0.061	-0.342	-0.029	0.043	0.259	0.312
Panel B: Bond Predictors							
Cochrane-Piazzesi factor	0.079	0.701	-3.630	-0.276	0.381	4.691	
Fama-Bliss spread, 5 Year	0.145	0.134	-0.374	0.047	0.251	0.423	
Fama-Bliss spread, 10 Year	0.183	0.160	-0.332	0.062	0.313	0.524	
Ludvigson-Ng factor	0.106	0.468	-1.919	-0.201	0.354	3.037	
Panel C: Stock Predictors							
Log dividend price ratio	-3.582	0.403	-4.524	-3.919	-3.310	-2.753	
Log dividend yield	-3.577	0.403	-4.531	-3.914	-3.306	-2.751	
Log earning price ratio	-2.825	0.439	-4.836	-2.993	-2.584	-1.899	
Log smooth earning price ratio	-3.075	0.341	-3.911	-3.275	-2.855	-2.274	
Log dividend payout ratio	-0.757	0.319	-1.244	-0.939	-0.601	1.379	
Book to market ratio	0.508	0.264	0.121	0.297	0.683	1.207	
T-Bill rate	0.049	0.031	0.000	0.030	0.064	0.163	
Long term yield	0.068	0.026	0.021	0.048	0.082	0.148	
Long term return	0.006	0.030	-0.112	-0.010	0.023	0.152	
Term spread	0.018	0.015	-0.036	0.007	0.031	0.045	
Default yield spread	0.010	0.005	0.003	0.007	0.012	0.034	
Default return spread	0.000	0.015	-0.098	-0.005	0.006	0.074	
Stock variance	0.002	0.004	0.000	0.001	0.002	0.065	
Net equity expansion	0.012	0.019	-0.058	0.003	0.025	0.051	
Inflation	0.003	0.003	-0.018	0.002	0.005	0.018	

This table provides summary statistics for the excess returns and predictors we consider in our analysis. Specifically, for each series we report the mean, standard deviation, minimum, quartiles, maximum, and, for the excess returns, the annualized Sharpe ratio. Panel A reports summary statistics for the stock and bond excess returns. All returns are continuously compounded. Monthly data on the stocks come from CRSP Cap-based portfolios file, where Large-cap stocks are the largest 20% of firms, Mid-cap are the 50th to 80th size percentile of firms, and Small-cap are the smallest half of firms. Monthly returns on five- and ten-year Treasury bonds are computed using the two-step procedure described in Gargano et al. (2017). Panel B reports summary statistics for the three bond predictors considered in Gargano et al. (2017). Panel C provides summary statistics for the 15 stock predictors considered in Welch and Goyal (2008). The sample period ranges from January 1962 to December 2014.

As for the SG-DLM, separately for each of the q equations in the system, we set the vectors $\boldsymbol{m}_{j0}$ in Equation (1.20) to the corresponding vectors of OLS estimates obtained over the training sample, while we set s_{j0} to the sample variance obtained from the OLS residuals ($j = 1, ..., q$). Next, we let $\boldsymbol{C}_{j0} = 100 s_{j0} I_{p_j+j-1}$, which renders the prior on $\boldsymbol{C}_{j0}$ uninformative and also guarantees that the implied prior moments on the initial SG-DLM regression parameters are equivalent to those from the W-DLM. Lastly, as with the W-DLM, we set the degrees of freedom n_{j0} to 10, effectively making the prior on σ_{j0}^{-2} flat and uninformative.

1.4 Empirical Results

In this section, we describe our empirical results. We will begin with an investigation of the role played by the various key features of our approach, with a particular emphasis on the importance of time variation in the first and second moments of asset returns and the strength of the predictability stemming from the various regressors we consider. Next, we will turn to examining the quality and accuracy of the W-DLM and SG-DLM forecasts, with an eye towards both statistical and economic measures of predictability. More specifically, we will evaluate the forecast accuracy of these models over the last 360 months of data in our sample, January 1985 through December 2014. In this way, we explicitly remove from the forecast evaluation sample the period of time characterized by the oil shocks of 1973-1974 and the bond market experimentation of the early 1980's.

1.4.1 A close look at the role of the various modeling features

As we discussed in Subsection 1.2.3, one of the key advantages of our approach is the ability to take into account the model uncertainty arising from both the availability of different predictor variables and the presence of multiple discount factors controlling the time variation in regression coefficients, variances, and covariances. Thanks to the closed-form

nature of the forward filters we rely on, this can be accomplished in a very timely manner, without the need to resort on expensive MCMC simulations. In this section, we take a close look at the role of predictor uncertainty and time variation in both the W-DLM and SG-DLM models.

In order to disentangle the relative importance of these features, for both the W-DLM and SG-DLM models we compute four variations of the score-weighted and equal-weighted model combinations described in Subsection 1.2.3. Our first model combination, which we label LIN, constrains $\delta_\beta = \delta_v = 1$ for the W-DLMs and $\delta_{\beta j} = \delta_{vj} = \delta_{\gamma j} = 1$ for the SG-DLMs ($j = 1, ..., q$), thus completely removing time variation in both the regression coefficients and variance-covariance matrix. In other words, the LIN specifications control for the uncertainty arising solely from the choice of which predictors to include in the model. Our second variant, which we denote as TVP to reference its time-varying parameters, is obtained by selectively combining only the subset of W-DLMs or SG-DLMs with $\delta_v = 1$ (as well as, in the case of the SG-DLMs, $\delta_\gamma = 1$) and $\delta_\beta < 1$. In this case, we are focusing on all those models with a constant variance covariance matrix, taking into account the uncertainty pertaining to the choice of which predictors to include and how much the regression coefficients should be allowed to vary over time.[21] Our third variant, which we denote as SV to reference stochastic volatility, is similarly obtained by selectively combining only the subset of models with $\delta_\beta = 1$ and $\delta_v < 1$ (as well as, in the case of the SG-DLMs, either $\delta_v < 1$ or $\delta_\gamma < 1$). Thus, we are removing altogether time variation in the regression coefficients, while controlling for the uncertainty arising from which predictors to include and how much time variation to afford to variances and covariances. Lastly, our fourth model combination variant, which we denote with TVP-SV, is obtained by combining all the W-DLMs or SG-DLMs that set $\delta_\beta < 1$ and $\delta_v < 1$ (as well as, in the case of the SG-DLMs, $\delta_\gamma < 1$). This is therefore our most flexible

[21] In this sense, we can think of the TVP variant of our model combinations as the multivariate extension of the approach first proposed by Dangl and Halling (2012) to forecast stock returns.

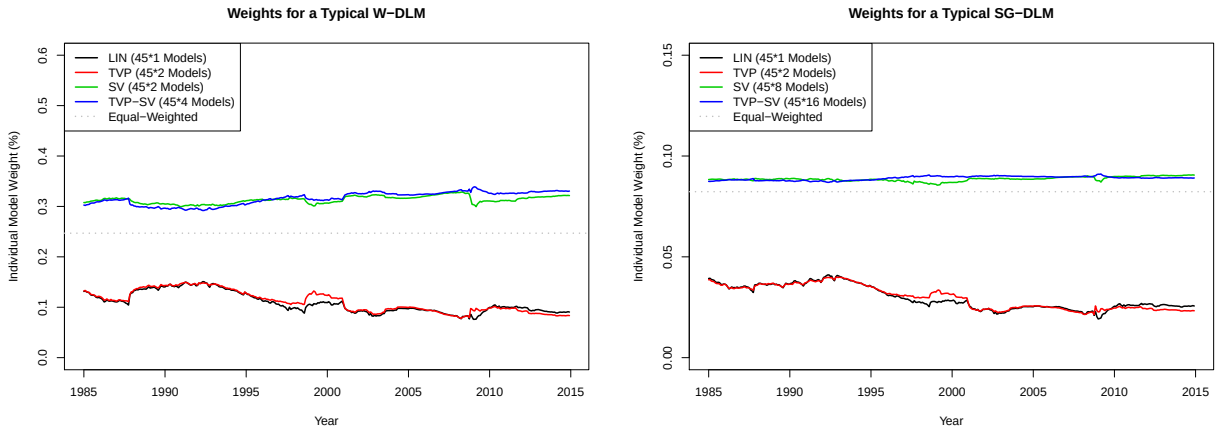
model combination scheme, where the regression coefficients, variances, and covariances all vary over time.

We start by looking at the importance of time variation in the first and second moments of stock and bond returns. The score-based model combination weights in Equation (1.32) are a combination of accuracy forecasting the first and second moments. To demonstrate how these weights change over time, instead of plotting curves for all 405 W-DLMs and 1215 SG-DLMs (as there are 45 different pairs of predictors matched with 9 and 27 different combinations of discount factors for the W-DLMs and SG-DLMs respectively), we summarize by plotting the percent of the total aggregate weight that a typical model from one of our model combinations receives. This is simply calculated as the percent of total model weight assigned to each one of the four groups (LIN/TVP/SV/TVP-SV) and dividing it by the total number of models within that group. Figure 1.1 shows the evolution over time of these weights for our four model combination variants of the W-DLMs (top panel) and SG-DLMs (bottom panel). We also report, in the legend of both panels, the total number of models within each group. As it can be seen from both panels, the assumption of constant variances/covariances appears to be strongly rejected by the data, with the average model combination weights of the models belonging to the LIN and TVP only receiving marginal support by the data. In contrast, allowing for variation in the variances and covariances produces much larger model weights, with the models within the SV and TVP-SV model combinations receiving, on average, weights that are two to three times larger.

Next, Figures 1.2 and 1.3 plot the time series of asset volatilities and cross-correlations associated with the score-weighted LIN, TVP, SV, and TVP-SV SG-DLM model specifications.[22] For comparability, we have also included in each panel the (constant) sample volatility or correlation, computed over the 1972–2014 period and depicted with a thin black line. For all five assets, we see that the conditional volatilities of all five assets for the SV

[22]We provide similar plots for the four W-DLM model combination variants in Appendix C.

Figure 1.1: Time series of score-based weights by feature set



This figure shows the evolution over time of the score-based model combination weights for the four variants of the W-DLM (top panel) and SG-DLM (bottom panel) models, namely LIN, TVP, SV, and TVP-SV. At each point in time t, we compute model $\mathcal{M}_i$'s weight ($i = 1, ..., K_W$ in the case of the W-DLM models and $i = 1, ..., K_{SG}$ in the case of the SG-DLM models) by looking at its historical statistical performance up through time $t-1$, as determined by the log score:

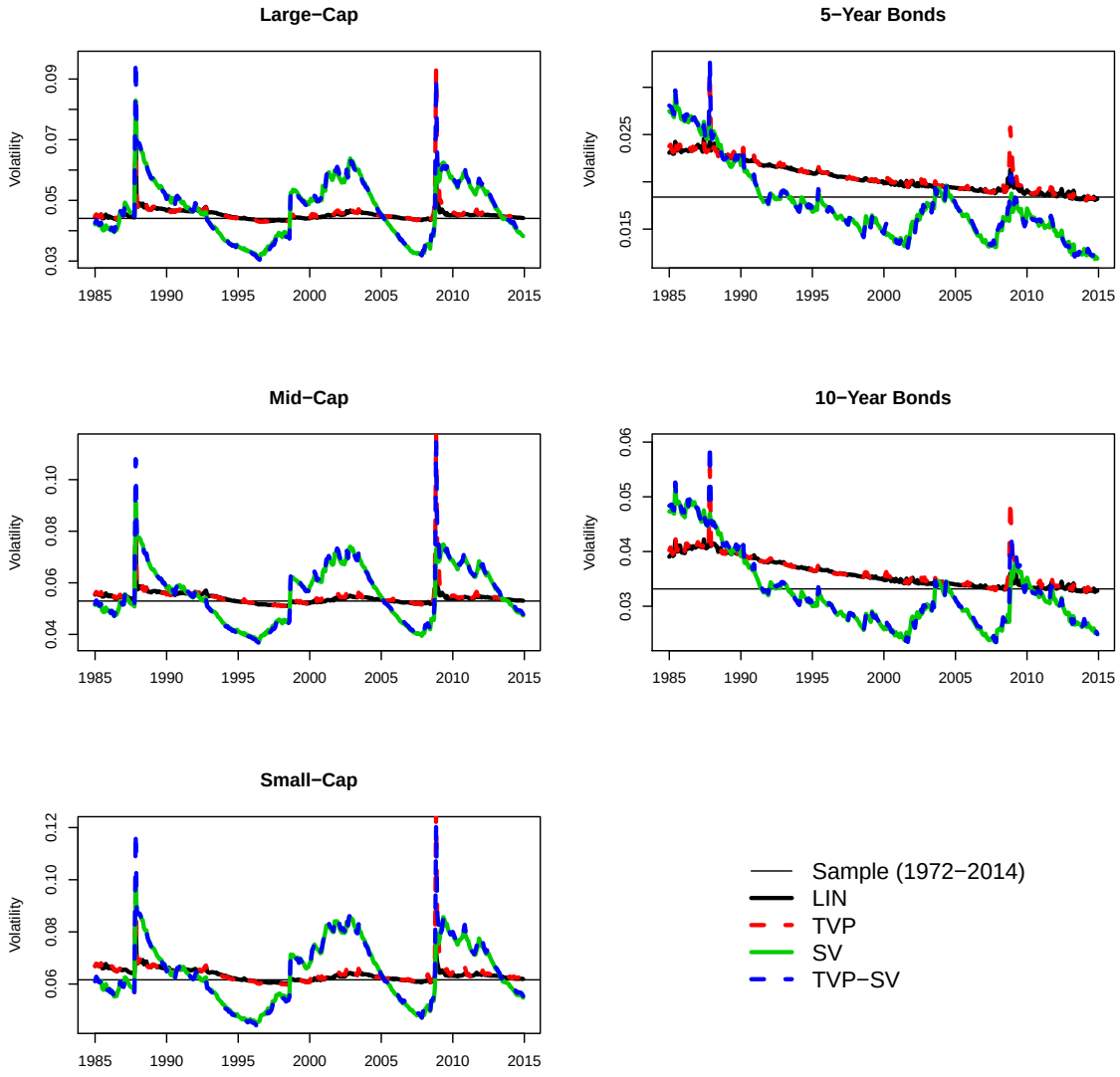
$$w_{i,t} \propto \sum_{\tau=1}^{t-1} \ln(S_{i,\tau})$$

where $S_{i,\tau}$ denotes the recursively computed score for model i at time τ, which we obtain by evaluating a Gaussian density with mean vector and covariance matrix equal to $\mathbb{E}(\boldsymbol{r}_\tau|\mathcal{M}_i, \mathcal{D}_{\tau-1})$ and $Cov(\boldsymbol{r}_\tau|\mathcal{M}_i, \mathcal{D}_{\tau-1})$ at the realized log excess returns $\boldsymbol{r}_\tau$. Next, we normalize the model weights across all models such that $\sum_{i=1}^{K_W} w_{i,t} = 1$ for W-DLMs and $\sum_{i=1}^{K_{SG}} w_{i,t} = 1$ for SG-DLMs. Then, models are aggregated to the appropriate feature set , whether it be LIN, TVP, SV, or TVP-SV. The curve shown here is the mean model weight as a percentage of the aggregate model weight over time, where the mean is taken over a given feature set. The evaluation period is January 1985 – December 2014.

and TVP-SV models vary significantly over time, with long spells of time characterized by above-average volatility (see in particular the late 1980's, early 2000's, and the most recent financial crisis) as well as shorter periods with below-average volatilities. We can also recognize a marked difference between the pattern of time variation of the three equity returns and those of the two bond returns, and strong similarities in the volatilities of the 5- and 10-year bond returns, which are seen in the different magnitudes of the vertical axes. As for the correlations, we see long stretches of time with conditional correlations that are either above or below their average counterparts. Again, we observe different patterns of time variation in the three equity returns from those of the two bond returns.

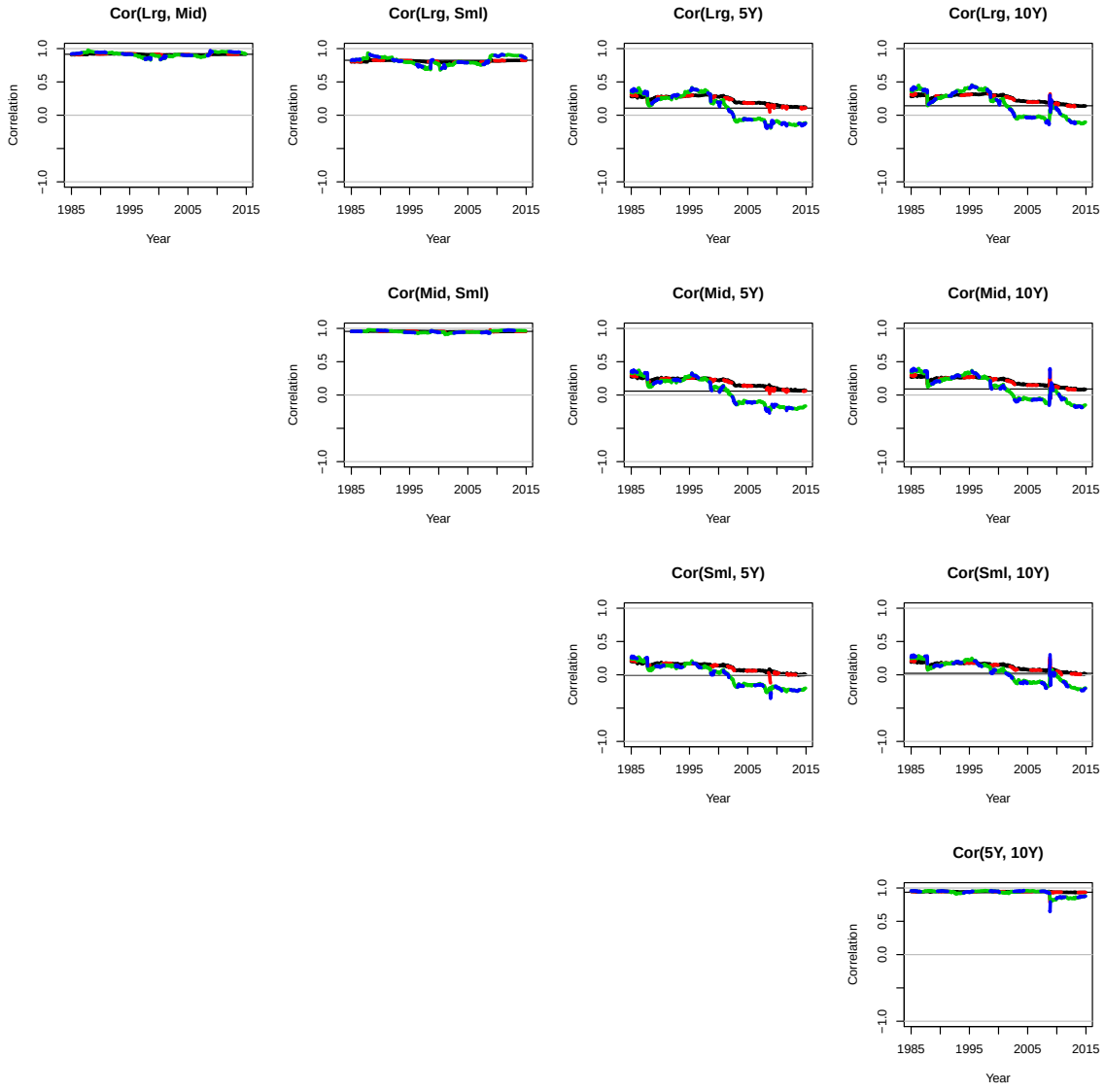
We conclude this section with a look at the relative importance of the predictor variables we consider in this study. Figure 1.4 depicts the evolution over time of the score-based model combination weights in Equation (1.32) for both the W- and SG-DLM over the evaluation period, January 1985 through December 2014. More specifically, the left panels of the figure focus on the equity predictors from Welch and Goyal (2008), and the right panels of the figure repeat the same calculations for the three bond predictors from Gargano et al. (2017). The construction of these curves is the same as in Figure 1.1, but instead of combining into different feature sets, weights are combined based on the predictors used. Starting with the left panels of the figure, we observe that among all the equity predictors, the stock variance, default yield spread, and default return spread take turns having the largest weight in the model combination, and are always among the most important variables. Conversely, the earning/price ratio predictor appears to not be favored in the model combination, consistently scoring among the lowest weights. Moving on to the right panel of the figure and the bond predictors, we find that the Cochrane-Piazzesi factor and the Fama-Bliss spreads receive the highest weights in the model combination, while surprisingly the Ludvigson-Ng macro factor appears to be less favored, at least in terms of log-scores. This result, which at first appears to be contradicting the results in Gargano et al. (2017), is due to the fact that

Figure 1.2: Time series of predicted volatilities for SG-DLM models



This figure shows the time-series of predicted volatilities of excess returns for the four variants of the SG-DLM score-based model combinations, namely LIN, TVP, SV, and TVP-SV. Each panel represents a different asset, as labeled. Note that the scales of the vertical axes are different for each asset in order to compare patterns of change over time, as opposed comparing the magnitude of volatilities across assets. The solid black line represents the LIN model; the dotted red line tracks the TVP model; the solid green line depicts the SV model, while the blue dotted line displays the TVP-SV model. In each panel we also display, as a reference, the level of the unconditional standard deviation of each asset, computed over the whole evaluation period, January 1972 – December 2014.

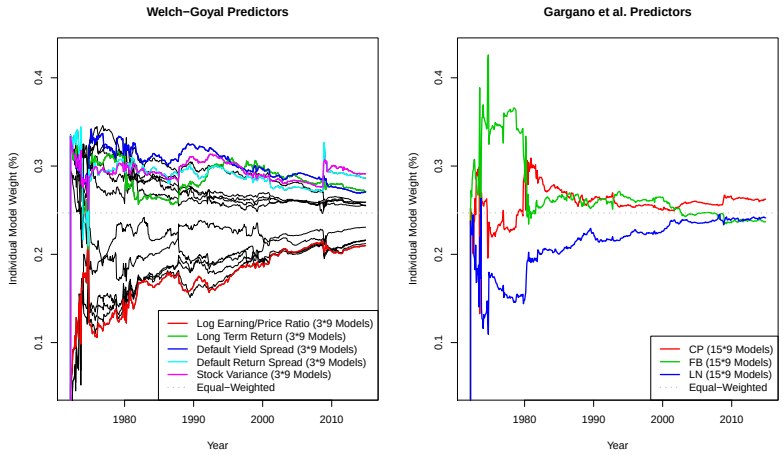
Figure 1.3: Time series of predicted correlations for SG-DLM models



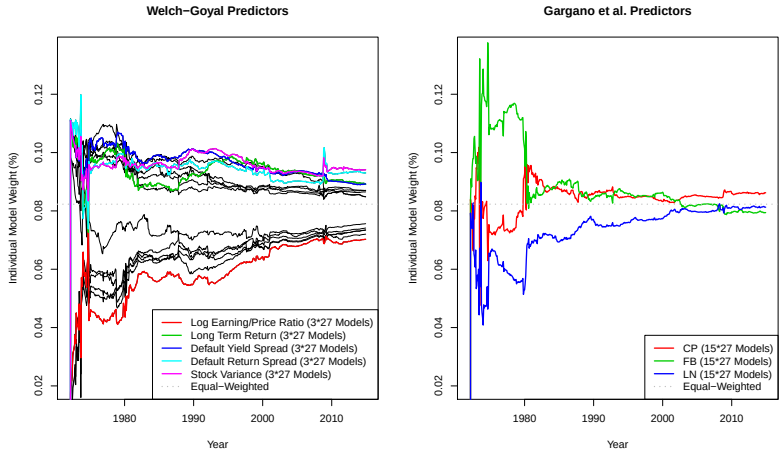
This figure shows the time-series of predicted correlations of excess returns for the four variants of the SG-DLM score-based model combinations, namely LIN, TVP, SV, and TVP-SV. Each panel represents a different pair of asset returns, as labeled. The solid black line represents the LIN model; the dotted red line tracks the TVP model; the solid green line depicts the SV model, while the blue dotted line displays the TVP-SV model. In each panel we also display, as a reference, the level of the unconditional correlation between each pair of asset returns, computed over the whole evaluation period, January 1972 – December 2014.

Figure 1.4: Time-series of score-based weights by predictor

Panel A: W-DLM



Panel B: SG-DLM



This figure shows the evolution over time of the score-based model combination weights for the various stock (left panels) and bond (right panels) predictors. The top two panels display results for the W-DLM models, while the bottom panels show the model combinations of the SG-DLM models. At each point in time t, we compute model $\mathcal{M}_i$ weight ($i = 1, ..., K_W$ in the case of the W-DLMs and $i = 1, ..., K_{SG}$ in the case of the SG-DLMs) by looking at its historical statistical performance up through time $t-1$, as determined by the log score:

$$w_{i,t} \propto \sum_{\tau=1}^{t-1} \ln(S_{i,\tau})$$

where $S_{i,\tau}$ denotes the recursively computed score for model i at time τ, which we obtain by evaluating a Gaussian density with mean vector and covariance matrix equal to $\mathbb{E}(\boldsymbol{r}_\tau|\mathcal{M}_i, \mathcal{D}_{\tau-1})$ and $Cov(\boldsymbol{r}_\tau|\mathcal{M}_i, \mathcal{D}_{\tau-1})$ at the realized log excess returns $\boldsymbol{r}_\tau$. Next, we normalize the model weights across all models such that $\sum_{i=1}^{K_W} w_{i,t} = 1$ for W-DLMs and $\sum_{i=1}^{K_{SG}} w_{i,t} = 1$ for SG-DLMs. Then, models are aggregated according to the predictors included in each model. The curve shown here is the mean model weight as a percentage of the aggregate model weight over time, where the mean is taken over all models with the listed predictor. The evaluation period is January 1985 – December 2014.

our combination weights are driven by the predictors' relative log-scores, and as Gargano et al. (2017, Figures 6-7) show, the advantage of the Ludvigson-Ng macro factor is particularly apparent when focusing on point predictability. However, it is worthwhile pointing out that the performance gap between the three predictors shrinks over time. The general patterns in the score-weights over time for W- and SG-DLMs are largely similar, showing that the relative importance of the variables holds regardless of DLM type.

1.4.2 Out-of-Sample Performance

We now turn to evaluating the relative predictive accuracy of the various W-DLM and SG-DLM specifications over the period spanning from January 1985 to December 2014. Throughout, our benchmark model will be a no-predictability SG-DLM with constant mean and constant variance-covariance matrix (that is, the specification in Equation (1.13) with $\boldsymbol{x}_{j,t-1} = 1$ and $\delta_{\beta_j} = \delta_{\gamma j} = \delta_{vj} = 1$, for all j), in line with what is customary in both the stock and bond return predictability literatures. We will provide results separately for each of the five risky assets that we are focusing on, as well jointly for the whole system of equations.

1.4.2.1 Measures of Predictive Accuracy

Starting with the point forecast accuracy, for each of the five asset returns we summarize the precision of the point forecasts for model i, relative to that from the benchmark model, by means of the ratio of mean-squared forecast errors ("MSFEs"):

$$MSFE_{ij} = \frac{\sum_{\tau=\underline{t}}^{T} e_{ij,\tau}^2}{\sum_{\tau=\underline{t}}^{T} e_{bcmk,j,\tau}^2} \tag{1.33}$$

where $\underline{t}$ denotes the beginning of the out-of-sample period, i refers to the W-DLM or SG-DLM model under consideration (i.e. LIN, TVP, SV, TVP-SV), $e_{ij,\tau} = r_{j\tau} - \mathbb{E}(r_{j\tau}|\mathcal{M}_i, \mathcal{D}_{\tau-1})$ and $e_{bcmk,j,\tau} = r_{j\tau} - \mathbb{E}(r_{j\tau}|\mathcal{M}_{bcmk}, \mathcal{D}_{\tau-1})$ are the forecast errors of asset return j at time τ associated with model i and the benchmark model, respectively. Values of $MSFE_{ij}$ below

one suggest that for asset j, model i produces more accurate point forecasts than the no-predictability benchmark.

We also measure the point-forecast accuracy of the various method by looking jointly at all five assets. Following Christoffersen and Diebold (1998), we compute the ratio between the weighted multivariate mean squared forecast error (WMSFE, also known as the squared Mahalanobis distance) of model i and the no-predictability benchmark as follows:

$$WMSFE_i = \frac{\sum_{\tau=\underline{t}}^{T} \boldsymbol{e}_{i\tau}' \left[\widehat{Cov(\boldsymbol{r}_t)}\right]^{-1} \boldsymbol{e}_{i\tau}}{\sum_{\tau=\underline{t}}^{T} \boldsymbol{e}_{bcmk,\tau}' \left[\widehat{Cov(\boldsymbol{r}_t)}\right]^{-1} \boldsymbol{e}_{bcmk,\tau}} \tag{1.34}$$

where $\boldsymbol{e}_{i\tau} = (e_{i1,\tau}, ..., e_{iq,\tau})'$ and $\boldsymbol{e}_{bcmk,\tau} = (e_{bcmk,1,\tau}, ..., e_{bcmk,q,\tau})'$ are the $q \times 1$ vectors of forecast errors at time τ associated with model i and the benchmark model, while $\widehat{Cov(\boldsymbol{r}_t)}$ denotes the sample estimates of the asset returns unconditional variance-covariance matrix, computed over the evaluation period.[23]

As for the quality of the density forecasts, we compute the average log score (ALS) differential between model i and the no-predictability benchmark,

$$ALS_{ij} = \frac{1}{T - \underline{t} + 1} \sum_{\tau=\underline{t}}^{T} \left(\ln\left(S_{ij,\tau}\right) - \ln\left(S_{bcmk,j,\tau}\right)\right) \tag{1.35}$$

where $S_{ij,\tau}$ ($S_{bcmk,j,\tau}$) denotes model i's (benchmark's) log score at time τ, which we obtain by evaluating a univariate Gaussian density with mean vector $\mathbb{E}(\boldsymbol{r}_{j\tau}|\mathcal{M}_i, \mathcal{D}_{\tau-1})$ and variance $Var(\boldsymbol{r}_{j\tau}|\mathcal{M}_i, \mathcal{D}_{\tau-1})$ (likewise $\mathbb{E}(\boldsymbol{r}_{j\tau}|\mathcal{M}_{bcmk}, \mathcal{D}_{\tau-1})$ and $Var(\boldsymbol{r}_{j\tau}|\mathcal{M}_{bcmk}, \mathcal{D}_{\tau-1})$) at the realized excess returns $\boldsymbol{r}_{j\tau}$. Positive values of ALS_{ij} indicate that model i produces on average

[23]The role of this covariance matrix is to standardize the distances in multivariate space, and weight the assets' forecast errors differently depending on the variability of the underlying assets and correlation across assets. All things equal, there will be less penalty for the forecast errors of a highly volatile asset than from those of a less-volatile asset, but also less reward when accurate. At the same time, there will be more penalty for forecast errors in directions not implied by the correlations in the empirical sample covariance matrix, meaning higher penalties for forecast errors in opposite directions for correlated assets and high penalties for forecast errors in the same direction for negatively correlated assets.

more accurate density forecasts for variable j than the benchmark. Finally, we consider the multivariate average log score differentials (MVALS) between model i and the benchmark,

$$MVALS_i = \frac{1}{T - \underline{t} + 1} \sum_{\tau=\underline{t}}^{T} \left(\ln \left(S_{i,\tau} \right) - \ln \left(S_{bcmk,\tau} \right) \right) \tag{1.36}$$

where $S_{i,\tau}$ ($S_{bcmk,\tau}$) are computed as described in Subsection 1.2.3.[24]

1.4.2.2 Results

We begin by inspecting the point and density forecast predictability of both W-DLM and SG-DLM models on an asset-by-asset basis, as summarized by the $MSFE$ and ALS metrics. Table 1.2 reports the $MSFE$ ratios of the LIN, SV, TVP, and TVP-SV variants of the W-DLM and SG-DLM models, individually for the five asset returns and relative to the no-predictability benchmark. Across the columns of the table, we report the average predictive improvements obtained by either relying on the equal-weighted or score-weighted combinations. As for the three equity returns, we see that stochastic volatility plays a small role in improving the SG-DLM predictions, while time-varying parameters hurt both W- and SG-DLMs. In fact, the TVP and TVP-SV specifications do worse at point prediction than the benchmark. Results for the two bond returns are stronger, with ample and widespread evidence of point-predictability. This appears to be true regardless of the combination scheme and the set of features considered, though time-varying parameters are again never preferable. This widespread predictability is consistent with the findings of both Thornton and Valente (2012) and Gargano et al. (2017). Next, Table 1.3 inspects the asset-specific density forecast predictability of the same models depicted in Table 1.2. Here, in line with the results reported in Figure 1.1, we find that for all five assets' SG-DLMs, the

[24]This measure penalizes wrong return predictions based on the variance of the prediction. If the model is highly confident in an inaccurate prediction, it scores very low. If highly confident and correct, it receives a high score. If the model is unconfident in the prediction, and hence has high variance and a relatively flat pdf, then there is little penalty for being wrong but also little bonus for being correct.

Table 1.2: Mean-squared forecast errors of W-DLM and SG-DLM models by asset

	W-DLM		SG-DLM	
Features	Equal	Score	Equal	Score
Panel A: Large-Cap				
LIN	0.998	0.998	0.998	0.997
TVP	1.008	1.007	1.009	1.007
SV	0.998	**0.997**	0.993	**0.993**
TVP-SV	1.008	1.007	1.008	1.007
Panel B: Mid-Cap				
LIN	**0.991**	0.993	0.991	0.993
TVP	1.015	1.013	1.016	1.014
SV	**0.991**	0.993	**0.986**	0.987
TVP-SV	1.015	1.013	1.015	1.014
Panel C: Small-Cap				
LIN	**0.989**	0.993	0.990	0.993
TVP	1.016	1.015	1.016	1.015
SV	**0.989**	0.993	**0.986**	0.987
TVP-SV	1.016	1.014	1.016	1.015
Panel D: 5-Year Bonds				
LIN	**0.962**	0.970	**0.962**	0.968
TVP	0.965	0.970	0.965	0.968
SV	**0.962**	0.964	**0.962**	0.963
TVP-SV	0.965	0.966	0.965	0.965
Panel E: 10-Year Bonds				
LIN	**0.965**	0.970	**0.965**	0.968
TVP	0.981	0.988	0.981	0.987
SV	**0.965**	0.966	0.967	0.967
TVP-SV	0.981	0.987	0.981	0.983

This table reports, for each of the five asset returns we considered, the ratio of mean-squared forecast errors ("MSFEs") between a given model and the no-predictability benchmark, computed as

$$MSFE_{ij} = \frac{\sum_{\tau=\underline{t}}^{T} e_{ij,\tau}^2}{\sum_{\tau=\underline{t}}^{T} e_{bcmk,j,\tau}^2}$$

where $\underline{t}$ denotes the beginning of the out-of-sample period, i refers to the model under consideration (i.e. LIN, TVP, SV, TVP-SV W-DLMs or SG-DLMs), $e_{ij,\tau} = r_{j\tau} - \mathbb{E}(r_{j\tau}|\mathcal{M}_i, \mathcal{D}_{\tau-1})$ and $e_{bcmk,j,\tau} = r_{j\tau} - \mathbb{E}(r_{j\tau}|\mathcal{M}_{bcmk}, \mathcal{D}_{\tau-1})$ are the forecast errors of asset return j at time τ associated with model i and the benchmark model, respectively. Values of $MSFE_{ij}$ below one suggest that for asset j, model i produces more accurate point forecasts than the no-predictability benchmark. Bold-faced values indicate the best performing models within each asset class and DLM type, while "Equal" and "Score" denote, respectively, equal-weighted and score-weighted model combinations. The evaluation period is January 1985 – December 2014.

SV and TVP-SV model combinations always lead to the most accurate predictive densities. However, we see this does not hold for W-DLMs on mid- and small-cap stocks, where LIN is best. Furthermore, the mid- and small-cap stocks see large drops in performance when adding time-varying parameters, which benefit 10-Year bonds, across models and weighting schemes.

Next, we turn to the joint point and density forecast predictability, as measured by the $WMSFE$ and $MVALS$ metrics. Starting with Table 1.4, we find that in term of point forecast accuracy the best performing models feature constant coefficients, and adding time-varying parameters to the model set appear to slightly increase the forecasting error, across the board. Moving on to the joint accuracy of the density forecasts, Table 1.5 reports the average (multivariate) log score improvements that are brought in by the different sets of feature and model combination schemes. The SV and TVP-SV feature sets lead to the largest gains in accuracy, with the largest gains being associated with the SG-DLM TVP-SV model. Interestingly, the comparison of log scores between W-DLM and SG-DLM seem to favor the latter model specification when volatilities and correlations are stochastic, while when constant, the DLM types are comparable. To shed further light on where the SV and TVP-SV model are most successful, Figure 1.5 plots the cumulative sum of the multivariate log score differentials, $CSMVLSD_{it} = \sum_{\tau=\underline{t}}^{t} \left(\ln\left(S_{i,\tau}\right) - \ln\left(S_{bcmk,\tau}\right)\right)$ for SG-DLMs with different feature sets. The figure clearly shows how, starting around the mid 1990's and continuing all the way to the end of the sample, the SV and TVP-SV models consistently generate significantly more accurate density forecasts than all the alternative model specifications. Furthermore, we note that TVP-SV gains a step over SV during the housing bubble, which step persists through the end of the sample.

The previous tables and figures indicate that time-varying volatility and correlation play a very important role in generating sharp density forecasts. This appears to be true both at the individual asset level as well as when focusing on all stock and bond returns

Table 1.3: Average log score differentials of W-DLM and SG-DLM models by asset

	W-DLM		SG-DLM	
Features	Equal	Score	Equal	Score
Panel A: Large-Cap				
LIN	0.007	0.006	0.006	0.006
TVP	0.006	0.005	0.006	0.005
SV	**0.016**	0.015	**0.012**	0.010
TVP-SV	0.014	0.014	0.010	0.007
Panel B: Mid-Cap				
LIN	**0.009**	0.007	0.008	0.007
TVP	0.002	0.002	0.002	0.001
SV	0.007	0.004	0.014	**0.015**
TVP-SV	-0.002	-0.002	0.006	0.006
Panel C: Small-Cap				
LIN	**0.010**	0.008	0.008	0.006
TVP	0.003	0.002	0.002	0.001
SV	0.005	0.001	0.017	**0.019**
TVP-SV	-0.006	-0.007	0.008	0.010
Panel D: 5-Year Bonds				
LIN	0.016	0.016	0.016	0.016
TVP	0.014	0.015	0.014	0.015
SV	**0.095**	0.093	0.078	**0.091**
TVP-SV	0.094	0.094	0.077	0.089
Panel E: 10-Year Bonds				
LIN	0.018	0.018	0.018	0.019
TVP	0.018	0.018	0.018	0.019
SV	0.062	0.061	0.058	0.070
TVP-SV	0.062	**0.063**	0.059	**0.070**

This table reports, for each of the five asset returns we considered, the average log score (ALS) differential between model i and the no-predictability benchmark,

$$ALS_{ij} = \frac{1}{T-\underline{t}+1}\sum_{\tau=\underline{t}}^{T}\left(\ln\left(S_{ij,\tau}\right) - \ln\left(S_{bcmk,j,\tau}\right)\right)$$

where $S_{ij,\tau}$ ($S_{bcmk,j,\tau}$) denotes model i's (benchmark's) log score at time τ, which we obtain by evaluating a univariate Gaussian density with mean vector $\mathbb{E}(\boldsymbol{r}_{j\tau}|\mathcal{M}_i,\mathcal{D}_{\tau-1})$ ($\mathbb{E}(\boldsymbol{r}_{j\tau}|\mathcal{M}_{bcmk},\mathcal{D}_{\tau-1})$) and variance $Var(\boldsymbol{r}_{j\tau}|\mathcal{M}_i,\mathcal{D}_{\tau-1})$ ($Var(\boldsymbol{r}_{j\tau}|\mathcal{M}_{bcmk},\mathcal{D}_{\tau-1})$) at the realized excess returns $\boldsymbol{r}_{j\tau}$. Positive values of ALS_{ij} indicate that model i produces on average more accurate density forecasts for variable j than the benchmark. Bold-faced values indicate the best performing models within each asset class and DLM type, while Equal and Score denote, respectively, equal-weighted and score-weighted model combinations. The evaluation period is January 1985 – December 2014.

Table 1.4: Weighted Mean-squared forecast errors of W-DLM and SG-DLM models

	W-DLM		SG-DLM	
Features, Weighting:	Equal	Score	Equal	Score
LIN	**0.989**	0.991	**0.989**	0.991
TVP	1.004	1.006	1.004	1.006
SV	**0.989**	0.990	0.990	0.990
TVP-SV	1.004	1.006	1.004	1.004

This table reports the ratio between the weighted multivariate mean squared forecast error (WMSFE, also known as the squared Mahalanobis distance) between model i and the no-predictability benchmark, computed as follows:

$$WMSFE_i = \frac{\sum_{\tau=\underline{t}}^{T} \boldsymbol{e}_{i\tau}' \left[\widehat{Cov(\boldsymbol{r}_t)}\right]^{-1} \boldsymbol{e}_{i\tau}}{\sum_{\tau=\underline{t}}^{T} \boldsymbol{e}_{bcmk,\tau}' \left[\widehat{Cov(\boldsymbol{r}_t)}\right]^{-1} \boldsymbol{e}_{bcmk,\tau}}$$

where $\boldsymbol{e}_{i\tau} = (e_{i1,\tau}, ..., e_{iq,\tau})'$ and $\boldsymbol{e}_{bcmk,\tau} = (e_{bcmk,1,\tau}, ..., e_{bcmk,q,\tau})'$ are the $q \times 1$ vector of forecast errors at time τ associated with model i and the benchmark model, while $\widehat{Cov(\boldsymbol{r}_t)}$ denotes the sample estimates of the asset returns unconditional variance-covariance matrix, computed over the evaluation period. Values of $WMSFE_i$ below one suggest that model i produces more accurate point forecasts than the no-predictability benchmark. Bold-faced values indicate the best performing models within DLM type, while Equal and Score denote, respectively, equal-weighted and score-weighted model combinations. The evaluation period is January 1985 – December 2014.

Table 1.5: Multivariate Average log score differentials of W-DLM and SG-DLM models

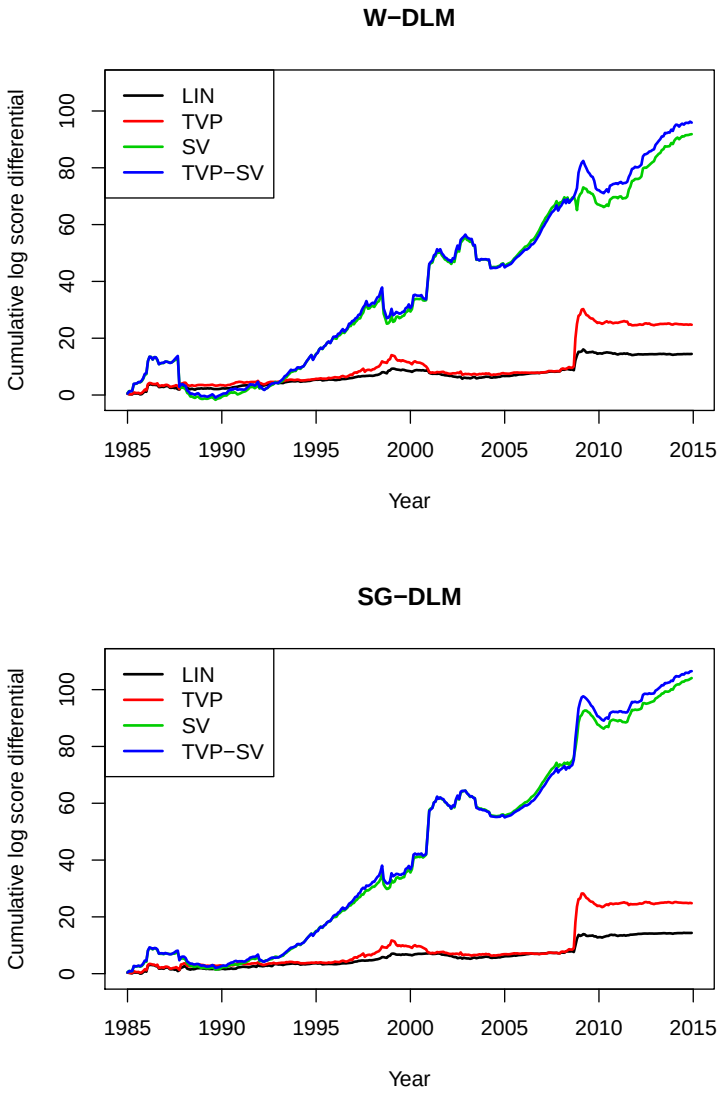
	W-DLM		SG-DLM	
Features, Weighting:	Equal	Score	Equal	Score
LIN	0.043	0.040	0.042	0.040
TVP	0.066	0.069	0.066	0.069
SV	0.266	0.255	0.271	0.289
TVP-SV	0.265	**0.266**	0.279	**0.296**

This table reports the multivariate average log score differentials (MVALS) between model i and the benchmark,

$$MVALS_i = \frac{1}{T - \underline{t} + 1} \sum_{\tau=\underline{t}}^{T} \left(\ln\left(S_{i,\tau}\right) - \ln\left(S_{bcmk,\tau}\right)\right)$$

where $S_{i,\tau}$ ($S_{bcmk,\tau}$) are computed as described in Subsection 1.2.3. Values of $MVALS_i$ above zero suggest that model i produces more accurate density forecasts than the no-predictability benchmark. Bold-faced values indicate the best performing models within DLM type, while Equal and Score denote, respectively, equal-weighted and score-weighted model combinations. The evaluation period is January 1985 – December 2014.

Figure 1.5: Cumulative sum of the multivariate log score differentials for W-DLM and SG-DLM models



This figure plots the cumulative sum of the multivariate log score differentials, $CSMVLSD_{it} = \sum_{\tau=\underline{t}}^{t} \left(\ln\left(S_{i,\tau}\right) - \ln\left(S_{bcmk,\tau}\right)\right)$ over time, where $S_{i,\tau}$ ($S_{bcmk,\tau}$) denote the model i's (benchmark's) log score and is computed as described in Subsection 1.2.3. The log score measures how accurate the multivariate distribution forecasts are given the realized observations. Within each panel, the solid black line represents the LIN model; the dotted red line tracks the TVP model; the solid green line depicts the SV model, while the blue dotted line displays the TVP-SV model. The evaluation period is January 1985 – December 2014.

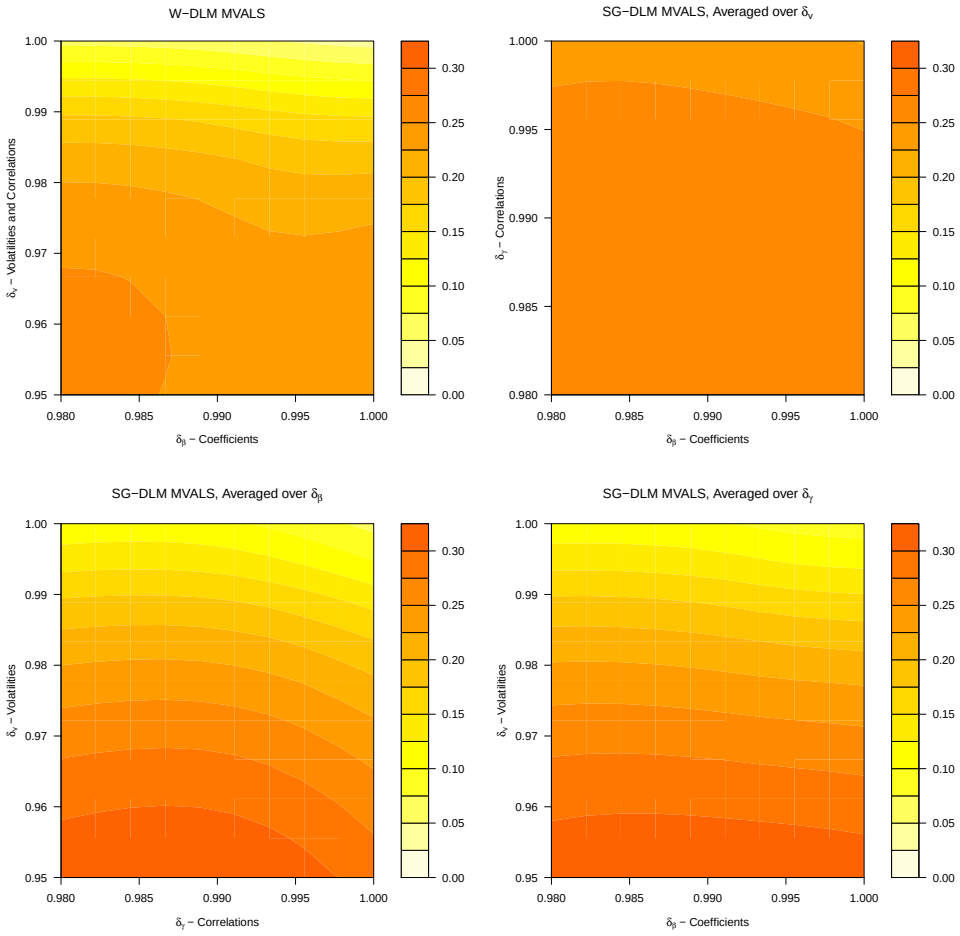
jointly. To offer additional insights on the mechanics behind this result, Figure 1.6 shows the heat-map of the joint density forecast accuracy, as measured by the $MVALS$ metric, for all the possible combinations of discount factors considered, both for the W-DLMs (top left panel) and the SG-DLMs (remaining panels).[25] This plot permits us to pinpoint exactly the mix of features that lead to the largest predictive gains in our model combinations, and to provide further clarity on the drivers of the results summarized in Tables 1.2 - 1.5. Starting with the W-DLM case, we observe that the most successful models feature a large degree of variation across the board, in volatilities, correlations, and regression coefficients. This demonstrates why there are high score values for the TVP-SV model in Table 1.5, as it is the only feature set incorporating models from this lower left quadrant. The remaining three panels show that the best performing SG-DLM models include high degrees of time variation in the volatilities, moderate degrees of time variation in correlations, and ambivalence toward the degree of movement in the regression coefficients (compared to the W-DLMs). Note specifically, however, that the ideal degree of time variation in the correlations ($\delta_\gamma \approx 0.986$) is much higher than that of volatility ($\delta_v \leq 0.95$). This further validates the importance of allowing for a different degree of variation in the on- and off-diagonal elements of the variance covariance matrix.

1.4.3 Portfolio Analysis

We now turn to evaluating the portfolio implications and economic predictability implied by the W-DLM and SG-DLM predictive densities that we derived.

[25] In particular, each point in the heat-maps corresponds to the average $MVALS$ associated to a given combination of discount factors, averaged over all 45 permutations of predictors (in the case of the SG-DLMs, also averaged over all possible values of the discount factor not shown on the axes).

Figure 1.6: Heat map of multivariate average log scores for different discount factors



This figure shows the multivariate average log scores ($MVALS$) of 100 different combinations of discount factors. The smaller a discount factor (δ) is, the more dynamic a feature is over time. δ_v controls the degree to which volatility is stochastic. δ_γ controls the degree to which correlations may time-vary. δ_β controls the time variation in the regression coefficients. In order to create a two-dimensional plot, each SG-DLM pane model-averages over one the three discount factors The evaluation period is January 1985 – December 2014.

1.4.3.1 Framework

We focus on the problem of a Bayesian investor endowed with power utility (see, among others, Johannes et al., 2014; Gargano et al., 2017; Gao and Nardari, 2018). At each point in time, the investor chooses her optimal asset allocations by distributing her total wealth between q (equity and bond) risky assets and one risk-free asset, under the constraint that the sum of her long and short positions does not exceed 300% of her wealth or fall below -200% of her wealth, and that none of her individual positions (including the weight on the risk-free asset) falls outside the same range.[26] Assuming that the excess returns on the q risky assets are jointly log-normally distributed, we can follow Campbell et al. (2003) and approximate $r_{p,it}$, the log return of the portfolio implied by model i at time t, with the following formula

$$r_{p,it} = r_{f,t-1} + \boldsymbol{\omega}'_{i,t-1}(\boldsymbol{r}_t - r_{f,t-1}\mathbf{1}) + \frac{1}{2}\boldsymbol{\omega}'_{i,t-1} diag(\widehat{\boldsymbol{\Sigma}}_{i,t|t-1}) - \frac{1}{2}\boldsymbol{\omega}'_{i,t-1}\widehat{\boldsymbol{\Sigma}}_{i,t|t-1}\boldsymbol{\omega}_{i,t-1} \tag{1.37}$$

where $\boldsymbol{\omega}_{i,t-1}$ is a vector of portfolio weights, $\widehat{\boldsymbol{\Sigma}}_{i,t|t-1} = Cov\left(\boldsymbol{r}_t | \mathcal{M}_i, \mathcal{D}_{t-1}\right)$ denotes the risky assets' forecasted variance-covariance matrix at time t based on the estimates given by model $\mathcal{M}_i$ and conditional on the information set at time $t-1$, $r_{f,t-1}$ represents the continuously compounded risk-free rate, and $\mathbf{1}$ is a vector of ones the same length as $\boldsymbol{r}_t$. Let A denote the investor's relative degree of risk aversion, then the optimal weights on the q risky assets implied by model i are given by the solution of the following constrained maximization problem,

$$\begin{aligned} \arg\max_{\boldsymbol{\omega}_{i,t-1}} \quad & \boldsymbol{\omega}'_{i,t-1}\left(\widehat{\boldsymbol{\mu}}_{i,t|t-1} + \frac{1}{2}diag(\widehat{\boldsymbol{\Sigma}}_{i,t|t-1})\right) - \frac{A}{2}\boldsymbol{\omega}'_{i,t-1}\widehat{\boldsymbol{\Sigma}}_{i,t|t-1}\boldsymbol{\omega}_{i,t-1} \\ s.t. \quad & \boldsymbol{\omega}'_{i,t-1}\mathbf{1} \in [-2, 3] \\ & \boldsymbol{\omega}_{ij,t-1} \in [-2, 3], \quad j = 1, ..., q \end{aligned} \tag{1.38}$$

[26] Similarly, Gao and Nardari (2018) consider an investor who is constrained and will not be allowed to short risky assets and/or borrow more than a certain amount of cash.

where $\widehat{\boldsymbol{\mu}}_{i,t|t-1} = \mathbb{E}(\boldsymbol{r}_t|\mathcal{M}_i, \mathcal{D}_{t-1})$ is the mean of the predictive density of the vector of risky assets $\boldsymbol{r}_t$, computed using the information set available at time $t-1$ and under model $\mathcal{M}_i$.

We next use the sequence of portfolio weights $\{\boldsymbol{\omega}_{i,t-1}\}_{t=\underline{t}}^{T}$ computed under the various W- and SG-DLM models as well as under the benchmark model to compute the investor's certainty equivalent returns (CERs), which can be further expressed in percentage annualized terms as:

$$CER_i = 100 \times \left(\left[\frac{1}{T-\underline{t}+1} \sum_{t=\underline{t}}^{T} \widehat{W}_{it}^{1-A} \right]^{\frac{12}{1-A}} - 1 \right) \tag{1.39}$$

where $\widehat{W}_{it} = \exp(r_{p,it})$ denotes the realized wealth at time t, as implied by model i.

1.4.3.2 Results

Table 1.6 presents the annualized CERs over the whole out-of-sample period for the various W-DLM and SG-DLM model combinations for an investor with power utility and coefficient of relative risk aversion $A = 5$. In particular, the table reports the CER gains relative to the no-predictability benchmark model, i.e. $CER_i - CER_{bcmk}$ (as a reference point, the annualized CER for the benchmark model over the same period is equal to 5.896%.) As it can be inferred from the table, all feature sets and all model averaging weighting schemes produce higher CERs than the no-predictability benchmark. This is true regardless of whether we focus on the W- or SG-DLM models. In addition, we make the following observations. First, as it was the case with the log score measures, the largest gains in CERs occur when volatility is allowed to vary over time. Across the board, the inclusion of stochastic volatilities and correlations always lead to larger CERs, and this is true whether or not one also allows for time variation in the regression coefficients. That is, SV produces higher CERs than LIN, and so does TVP-SV compared to TVP. In particular, we find that the SV model combination of SG-DLMs produces CER gains of about 5.9% over the benchmark. The role of SV in the case of the W-DLM models is slightly less pronounced, with

Table 1.6: Annualized certainty equivalent returns of W-DLM and SG-DLM models

	W-DLM		SG-DLM	
Features, Weighting:	Equal	Score	Equal	Score
LIN	4.608	3.804	4.526	3.814
TVP	0.361	0.262	0.513	0.460
SV	**5.429**	4.623	**5.944**	5.892
TVP-SV	2.940	2.628	3.772	3.947

This table reports the annualized CERs (in percentage terms) over the whole out-of-sample period for the various W-DLM and SG-DLM model combinations for an investor with power utility and coefficient of relative risk aversion $A = 5$. In particular, the table reports the CER gains relative to the no-predictability benchmark model, i.e. $CER_i - CER_{bcmk}$, where

$$CER_i = 100 \times \left(\left[\frac{1}{T - \underline{t} + 1} \sum_{t=\underline{t}}^{T} \widehat{W}_{it}^{1-A} \right]^{\frac{12}{1-A}} - 1 \right)$$

and where $\widehat{W}_{it} = \exp(r_{p,it})$ denotes the realized wealth at time t, as implied by model i (the CER of the benchmark model, 5.896, is computed in an analogous manner). Bold-faced values indicate the best performing models within DLM type, while Equal and Score denote, respectively, equal-weighted and score-weighted model combinations. The evaluation period is January 1985 – December 2014.

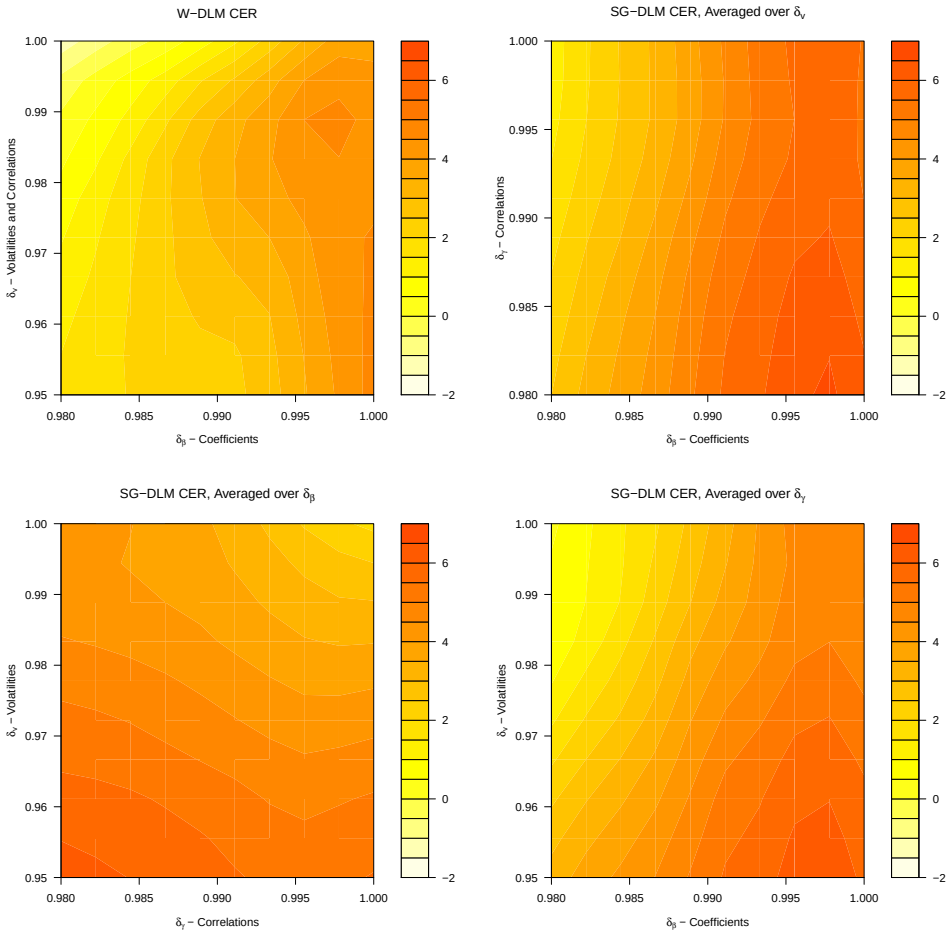
the improvements ranging from 4.6% to 5.4%. We attribute this difference to the additional restrictions that the W-DLM models impose, most notably the requirement that a single discount factor must control the degree of time variation in both variances and covariances. Second, the inclusion of time-varying coefficients in the model always decreases CERs. Third, equal-weighting holds a slight edge over the comparable score-weighted models, except for TVP-SV SG-DLMs. While the differences between the two weighting schemes are only marginal, this result goes in the opposite direction of what we found in Table 1.5 when focusing on the log score differentials, and is likely due to the way we computed the model combination weights in Equation (1.32). Fourth, SG-DLMs improve over the comparable W-DLMs, except for the equal-weighting LIN specification. Again, we believe this result is particularly due to the added flexibility that the SG-DLMs bring to the SV dynamics, and this is consistent with the largest differences between W- and SG-DLMs occurring when the SV feature is included in the model set.

As with our investigation of the joint statistical predictability of the W- and SG-DLM models, we conclude this section with a closer look at how exactly the various model features help achieve large CERs. In particular, Figure 1.7 shows the heat-map of the CER gains associated with all the possible combinations of discount factors considered, both for the W-DLMs (top left panel) and the SG-DLMs (remaining panels). Starting with the W-DLM case, we observe that there are two regions of highly profitable models. The first features a modest degree of time variation in the volatilities and correlations and very little variation in the regression coefficients. The second features no time variation in the regression coefficients to go along with high degrees of time variation in the covariance matrix elements. As for the more flexible SG-DLMs, the remaining three panels show that the best performing models also feature minimal (but some) variation in the regression coefficients, combined with as much time variation in the volatilities and correlations as we will allow. While the differences between W- and SG-DLM's optimal combinations for CER performance are not as stark as those for score, we do see that the magnitude of the SG-DLM CER noticeably exceed those of the W-DLM.

1.5 Conclusion

In this paper, we build on the Wishart dynamic linear model (W-DLM) of West and Harrison (1997) and the simultaneous graphical dynamic linear model (SG-DLM) of Gruber and West (2016) to introduce a flexible approach to model and forecast multiple asset returns. This approach allows us to integrate a number of useful features into a predictive system, namely model and parameter uncertainty, time-varying parameters, stochastic volatility, and time-varying covariances. We combine these DLM methods with a fully automated data-based model-averaging procedure to objectively determine the optimal set of said features and employ it to jointly forecast monthly stock and bond excess returns. This is made possible by the computational speed of the DLMs.

Figure 1.7: Heat map of certainty equivalent returns for different discount factors



This figure shows the certainty equivalent return differentials of 100 different combinations of discount factors. The smaller a discount factor (δ) is, the more dynamic a feature is over time. δ_v controls the degree to which volatility is stochastic. δ_γ controls the degree to which correlations may time-vary. δ_β controls the time variation in the regression coefficient parameters. In order to create a two-dimensional plot, each SG-DLM pane model-averages over one the three discount factors. The evaluation period runs from 1985 through 2014.

When evaluated over the January 1985 – December 2014 period, we find large statistical and economic benefits from using the appropriate ensemble of features in predicting stock and bond returns. In particular, we find that W-DLMs and SG-DLMs with stochastic volatility and time-varying covariances bring the largest gains in terms of statistical predictability, and that time-varying parameters can enhance the ensemble when forecasting distributions, though not for point predictions. Lastly, SG-DLM models with predictors, stochastic volatility and time-varying correlations lead to the largest economic gains. We show that when using this optimal set of features, a leverage-constrained power utility investor earns over 500 basis points (on an annualized basis) more than if she relied on the no-predictability benchmark.

Chapter 2

Monotonic Effects of Characteristics on Returns

This chapter is based on the text and content in Fisher et al. (2019b). We present a decision-theoretic method of choosing which firm characteristics are needed to explain the cross section of returns. This method requires a model, and this is an additive model of highly flexible splines. We balance this flexibility by examining how much structure, in the form of monotonic constraints, should be placed on characteristics' relationships with excess returns. Furthermore, shrinkage priors are placed on the spline coefficients.

2.1 Introduction

This paper considers the problem of predicting a firm's stock return with observable firm characteristics. These characteristics may be accounting measures such as market capitalization and book value as well as other observables such as a firm's past performance. Let r_{it} be the excess return of firm i at time t and let characteristics be incorporated into the vector $\boldsymbol{x}_{i,t-1}$. The conditional mean function

$$\mathbb{E}(r_{it} \mid \boldsymbol{x}_{i,t-1}) = f(\boldsymbol{x}_{i,t-1}) \tag{2.1}$$

is the object of interest. This paper accomplishes two goals. First, we develop a flexible Bayesian model for f. We carefully examine the statistical benefits of theoretically-motivated monotonicity constraints and time variation for our case study. These model features are previously unexplored in the finance literature, and we adapt methods from Shively et al. (2009) and McCarthy and Jensen (2016) to accomplish this goal. Second, we present a decision-theoretic framework for identifying the most predictive characteristics within the

vector $\boldsymbol{x}_{i,t-1}$, extending recent work in utility-based posterior summarization Hahn and Carvalho (2015); Puelz et al. (2015, 2017b,a, 2018) to nonlinear models.

Discovery of monotonic relationships in finance began decades ago. Fama and French (1993a) found that, on average, smaller firms have higher returns than larger firms. Jegadeesh and Titman (1993, 2001) documented that, on average, previously well performing firms (past winners) continue to do well in the near future, and past losers have low returns in the future. Patton and Timmermann (2010) develop statistical tests for monotonicity in assets returns. However, work still remains to understand the modeling impact of these revelations. In statistics, incorporating monotonicity constraints into models is a known but underutilized tool; see Shively et al. (2009) and Chipman et al. (2016) for recent developments. This paper presents a case study that combines decades-old empirical beliefs of monotonicity with this exciting and new statistical modeling work.

The list of potentially predictive characteristics is long and continues to grow, and numerous studies in finance have shown that these characteristics are independently useful for modeling returns. Harvey et al. (2016) catalog over 300 such characteristics and factors. Recently termed the "Factor Zoo" by Cochrane (2011) due to the sheer number of proposed characteristics and factors, the presence of hundreds is misleading, however, as many characteristics are likely drawing information from the same latent attributes of these firms and the economy. Understanding f as well as its characteristic-inputs is a venerable and urgent case study in finance and asset pricing. Hence, this paper will address the following questions: *Which characteristics are important?* And furthermore: *When are these characteristics important?* Of course, the size of $\boldsymbol{x}_{i,t-1}$ and the stationarity of the relationships between characteristics and returns depend on the choice of f, which brings us to a third question: *What do these relationships look like?*

2.1.1 Literature and Contributions

A traditional approach for understanding f is modeling the cross-section of firm returns as linearly related to a set of firm characteristics. Finance data, especially company return data, is a low-signal, high-noise environment, and structure is helpful to deal with this tremendous noise. Linear regression represents one extreme of model structure and simplicity, and most papers have at least some regression analysis, most famously the methods presented in Fama and MacBeth (1973). These methods are popular not only because of their structure but also interpretability. Linear regression is widely-known, easily estimated, and returns a single number representing the relationship between X and Y (the slope). Yet, as Freyberger et al. (2019) state, "no *a priori* reason exists why the conditional mean function should be linear." The core assumption of this standard approach may not hold. Therefore, recent literature considers nonlinear models for return-characteristic relationship (Freyberger et al., 2019; Gu et al., 2018). These methods utilize machine learning (ML) methods to infer nonlinear and joint relationships among characteristics and lie at the other modeling extreme: highly flexible but minimally interpretable. In this paper, we show that ML methods provide surprisingly poor predictive ability in the application area of finance and especially asset pricing.

An alternative, nonparametric, nonlinear approach for modeling f that maintains interpretability is *portfolio sorting*. This is done by cross-sectionally ranking firms based on an explanatory variable and computing the average firm return within each decile (or other quantile). Cattaneo et al. (2019) and Freyberger et al. (2019) show that this approach is essentially fitting a step function to return-characteristic relationship, as opposed to a linear fit typically from regression. While not always thought of as our conditional expectation function f, practitioners are computing such a function while calculating the average return for each decile of a predictive covariate. However, a step function is a simplistic functional form of the the return-characteristic relationship, as it must be assumed constant within

deciles and no information is shared across deciles. Fama and French (2008) summarize these issues in saying "sorts are clumsy for examining the functional form of the relation between average returns and an anomaly variable." Additionally, one quickly encounters dimensionality issues. Fitting a mean to each sorted decile of p variables requires 10^p datapoints, which very quickly is not plausible. Furthermore, Cattaneo et al. (2019) show that using 10 portfolios (deciles) is not enough, and that it is optimal to use more.

The methodology presented in this paper is most similar to Freyberger et al. (2019). We model f using additive quadratic splines, and this provides interpretability and flexibility. Our paper differs from Freyberger et al. (2019) in four important ways: We (i) characterize uncertainty through a fully Bayesian framework, (ii) examine the theoretical and statistical benefits of monotonicity constraints incorporated through priors, (iii) account for time variation through a first-principled, power-weighting density approach, and (iv) utilize statistical uncertainty to select the meaningful characteristics at each point in time. Standard unexplained volume, short-term reversal, market capitalization (size), and variants of momentum are found to be significant characteristics, and there is evidence this set changes in time. The data also provide strong support for monotonicity and time variability of the expected return function.

The rest of the paper proceeds as follows. Section 2.2 details the modeling methodology which relates to contributions (i)-(iii) above. Section 2.3 presents a simulation study to describe the merit of using monotonicity for structure and using power-weighting densities for nonstationarity. Section 2.4 details the utility-based posterior summarization that is used to select the meaningful characteristics which corresponds to contribution (iv) above. Section 2.5 reports the results from both the modeling and selection processes. Section 2.6 concludes.

2.2 Modeling Methodology

As discussed in Section 2.1, our model is comprised of the following components motivated by the case study: interpretability through additivity, flexibility through nonlinearity, minimal/specific structure through monotonicity, uncertainty through Bayesian priors, and nonstationarity through weighted densities. We outline each component in detail below.

Interpretability via an additive model. We address the first modeling objective by using an additive model, such that each characteristic's effect is separable from the others. Let

$$\mathbb{E}(r_{it}|\boldsymbol{x}_{i,t-1}) = \alpha_t + \sum_{k=1}^{K} f_{kt}(x_{i,k,t-1}) \tag{2.2}$$

where r_{it} is the time t return for firm i, α_t is the intercept term for time t, and $\boldsymbol{x}_{i,t-1}$ is a K length vector of firm i's characteristics at time $t-1$, where each characteristic is individually ranked across all n_{t-1} firms at time $t-1$

$$x_{i,k,t-1} = \frac{rank_{k,t-1}(characteristic_{i,k,t-1})}{n_t + 1}. \tag{2.3}$$

Thus, $x_{i,k,t-1} \in (0,1)$ is the empirical quantile of characteristic k for firm i at time $t-1$. This rank transformation is done to eliminate two issues with the predictors variables: (i) outliers and (ii) changes in the range of characteristics over time. For example, the market capitalization (size) of firms in general has increased, and a one billion dollar firm today might be in the 10th percentile of size while 30 years ago it was in the 90th percentile. Using the "empirical percentiles" from the rank transformation eliminates these issues as we only look at a firm's relative place in the distribution of a given characteristic; Freyberger et al. (2019) scale characteristics in the same way.

However, we propose a novel adjustment. The intercept in Equation (2.2) is the expected return when all x's are 0, and under the rank transformation $x_{i,k,t-1} = 0$, $\forall k$, means the smallest possible value for x, across all variables. The intercept α_t in Equation

(2.2) would be interpreted as the average return for a "perfectly minimum" firm, that is, a firm with the lowest value of each characteristic across all firms. This firm does not reasonably exist. As such, we shift the x-space by setting

$$x_{i,k,t-1} = \frac{rank_{k,t-1}(characteristic_{i,k,t-1})}{n_t + 1} - .5 \tag{2.4}$$

such that $x_{i,k,t-1} \in (-0.5, 0.5)$. Now, the intercept α_t represents the average return for a "perfectly median" firm, that is, a firm that has the median value across all characteristics.

Nonlinearity through quadratic splines. We address the second modeling objective through the use of quadratic splines. Typically, this would mean

$$f(x) = \beta_1 x + \beta_2 (x)^2 + \beta_3 (x - \acute{x}_1)_+^2 + ... + \beta_{\acute{m}+2} (x - \acute{x}_m)_+^2 \tag{2.5}$$

for m knots, $0 < \acute{x}_1 < ... < \acute{x}_m < 1$, where $(y)_+ = max(0, y)$.

However, our intercept adjustment requires an adjustment to the standard notation. Let f_{kt} be the quadratic spline for characteristic k at time t. For now, we'll drop the ikt subscripts for simplicity. For a given series of $\grave{m}+1$ nonpositive knots ($\grave{x}_{\grave{m}} < ... < \grave{x}_1 < \grave{x}_0 = 0$) and $\acute{m} + 1$ nonnegative knots ($0 = \acute{x}_0 < \acute{x}_1 < ... < \acute{x}_{\acute{m}}$), we set

$$\begin{aligned} f(x) =& \beta_1 x + \beta_2 (x)_-^2 + \beta_3 (x - \grave{x}_1)_-^2 + ... + \beta_{\grave{m}+2} (x - \grave{x}_{\grave{m}})_-^2 && (2.6) \\ &+ \beta_{\grave{m}+3} (x)_+^2 + \beta_{\grave{m}+4} (x - \acute{x}_1)_+^2 + ... + \beta_{\grave{m}+\acute{m}+3} (x - \acute{x}_{\acute{m}})_+^2 && (2.7) \end{aligned}$$

where the $(y)_+ = max(0, y)$ and $(y)_- = min(0, y)$. This can be abbreviated as $f(x) = \boldsymbol{x}^{*\prime}\boldsymbol{\beta}$ where $\boldsymbol{x}^*$ is the carefully constructed quadratic spline basis.

Structure imposed through monotonicity. Theoretical or *a priori* information can be used to add structure to these splines. We implement this through monotonicity constraints. Without loss of generality, we create these splines to be nondecreasing (can be nonincreasing)

using the ideas of Shively et al. (2009), Section 3, adapted to have both positive and negative knots.

By definition, the spline is monotonic nondecreasing if the first derivative is nonnegative for all x: $f'(x) \geq 0$. While specifications are in Appendix E, we suffice it here to say that the above restriction yields $\grave{m} + \acute{m} + 3$ linear constraints to satisfy, which can be summarized in a lower triangular matrix. We label this matrix $\boldsymbol{L}$ such that $\mathbf{0} \leq \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{\gamma}$, and we see that $\boldsymbol{L}$ acts as a projection matrix, projecting our more complicated constraints on $\boldsymbol{\beta}$ to the simple nonnegative constraints on $\boldsymbol{\gamma}$. Hence

$$f(x) = \boldsymbol{x}^{*\prime}\boldsymbol{\beta} = \boldsymbol{x}^{*\prime}L^{-1}L\boldsymbol{\beta} = \boldsymbol{w}'\boldsymbol{\gamma} \tag{2.8}$$

where $\boldsymbol{w}' = \boldsymbol{x}^{*\prime}L^{-1}$ is now our modified spline basis. Returning the use of subscripts ikt, Equation (2.2) is now

$$\mathbb{E}(r_{it}|\boldsymbol{x}_{i,t-1}) = \alpha_t + \sum_{k=1}^{K} \boldsymbol{w}'_{ikt}\boldsymbol{\gamma}_{kt}. \tag{2.9}$$

We allow our splines to be monotonic if there is prior information about the direction of a relationship between a firm characteristic and its stock return. For example, if we believe that a smaller firm will, on average, have higher returns than a larger firm, regardless of their absolute size Fama and French (1993a), then we believe size should have a monotonic relationship with expected returns. Monotonicity is one of the less intrusive structures we can assume to reign in the flexibility, and potential overfit, of splines. We demonstrate that enforcing monotonicity has statistical benefits as well as a useful interpretation. When the data is especially noisy, monotonicity is helpful in decreasing the variability of the inferred relationship between stock returns and characteristics.

Bayesian model specification. With Equation (2.9) introduced, we can describe the statistical model on our uncertainty. Let

$$r_{it} = \alpha_t + \sum_{k=1}^{K} \boldsymbol{w}'_{ikt}\boldsymbol{\gamma}_{kt} + \epsilon_{it} \tag{2.10}$$

with $\epsilon_{it} \sim N(0, \sigma^2)$.

We now set a prior on the coefficients γ. To protect against the overspecifying the number of knots, we include shrinkage as an important part of this prior. Let $I_{jkt} = 1$ indicate that $\gamma_{jkt} > 0$ and $I_{jkt} = 0$ indicate that $\gamma_{jkt} = 0$, where j indexes the $\grave{m} + \acute{m} + 3$ coefficients. Thus, I_{jkt} is a Bernoulli random variable with prior probability $P(I_{jkt} = 1) = p_{jk}$. This leads us to the conditional prior on γ_{jkt}:

$$(\gamma_{kj}|I_{kj} = 1, \cdot) \sim N_+(0, c_k\sigma^2) \tag{2.11}$$

where N_+ indicates a truncated Normal distribution with support on positive numbers (to change this entire setup to monotonic decreasing splines, we would simply change the support to negative numbers and appropriately adjust the definition of I_{jkt} above).

This setup allows us to let the data select the knots for the splines. By over-specifying the number of potential knots, the data will inform the model as to which knots should be included ($I_{jkt} = 1$) and which should not ($I_{jkt} = 0$).

Following Shively et al. (2009), we place uninformative priors on $\sigma^2 \sim U(0, 10^3)$ and $\alpha \sim N(0, 10^{10})$, as well as set $p_{jk} = 0.2$, $\forall j, k$. c_k is chosen, $\forall k$, to be 2253.689, the average number of firms in a quarter, across all quarters.

Nonstationarity incorporated through power-weighted densities. While using all historic data (i.e. using all data up to and including time $t-1$ to forecast time t events) is an option, this does not allow the parameters to adjust to trends over time (nonstationarity). Hence, we look at two approaches. First, we look at the traditional rolling-window method, where a model uses the most recent M time periods only, dropping all time periods older than the cutoffs. In this paper, we use $M = 120$ months, akin to much of the empirical finance literature.

Second, we use the power-weighted likelihood approach of McCarthy and Jensen

(2016). For $\omega_t \in [0,1]$, such that $\omega_1 \leq \omega_2 \leq ... \leq \omega_\tau$, the likelihood at time $\tau \in \{1,...,T\}$ discounts the impact of past data: $p(\boldsymbol{r}_1,...,\boldsymbol{r}_\tau|\Theta_\tau) = \prod_{t=1}^{\tau} p(\boldsymbol{r}_t|\Theta_\tau)^{\omega_t}$, to allow more recent data to receive more weight than older data, we choose $\omega_t = \delta^{\tau-t}$, for $\delta \in (0,1]$. Hence, for $\delta = 0.99$, yesterday's ω is 99% of today's. Thus, these likelihoods have an asymptotic effective sample size of $\frac{1}{1-\delta}$, e.g. $\frac{1}{1-0.99} = 100$.

McCarthy and Jensen (2016) point out, this is a simpler alternative to specifying a model for the evolution process itself. They also point out that the rolling window method is a special case of these power weights, such that $\omega_1 = ... = \omega_{\tau-120} = 0$ and $\omega_{\tau-119} = ... = \omega_\tau = 1$.

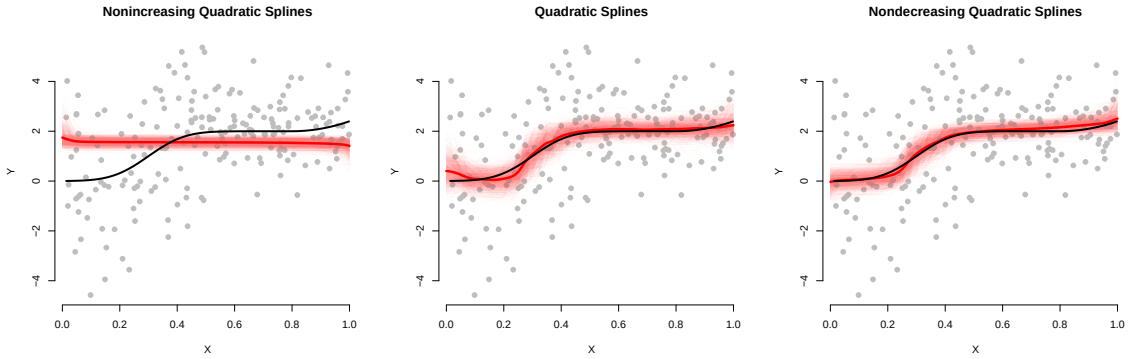
2.3 Simulation

Why monotonicity? When modeling functional phenomena, if the underlying generative function is in fact monotonic, then assuming monotonicity will improve your model. Specifically, the uncertainty about your fitted curve will be smaller, or in other words, the posterior will be more precise.

In Figure 2.1, we present a monotonic increasing mean function. The gray data points are randomly generated with heteroskedastic noise. Here, we model the data using varying monotonicity constraints. Posterior curve draws are shown in pink, and the posterior mean curve is in red. We see that while the unconstrained quadratic spline fits the underlying function reasonably well, the monotonic constrained spline fits better. Lastly, enforcing inappropriate constraints, namely a nonincreasing constraint in this case, disables the model. Hence, adding wise constraints help models ignore more of the noise and better detect signal.

We believe there is signal in the firm characteristics data we're analyzing, but there is a lot of noise, so this property of the model is desirable. When there is a weak signal (a barely-nonzero generative function), but low noise, models perform about equally with and without monotonicity. Again, we show in the top row of Figure 2.2 where the generative

Figure 2.1: Fits to data simulated from underlying monotonic function



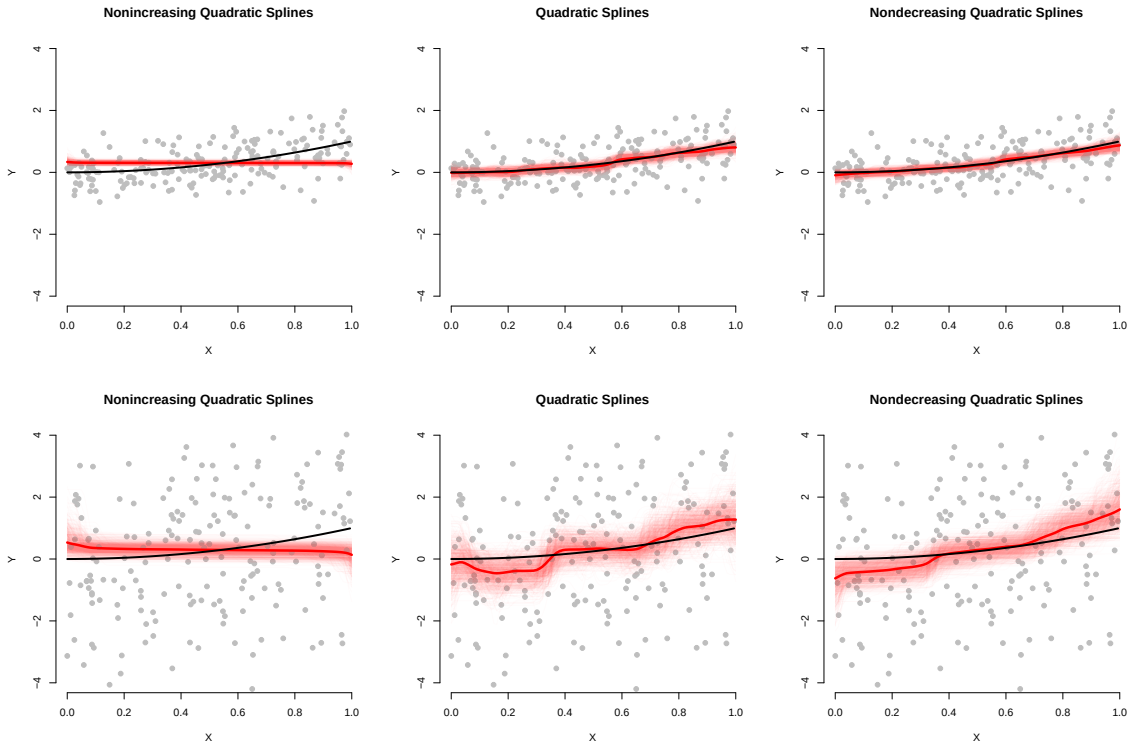
Data generated from a monotonically increasing mean function. Shown are three spline fits to the simulated data: (left) nonincreasing quadratic splines, (middle) quadratic splines without constraints, and (right) nondecreasing quadratic splines.

curve in black, data generated with homoskedastic noise in gray, the posterior draws in pink, and the posterior mean curve in red. However, as noise increases, the unconstrained spline tends to overfit to the data as in the bottom row of Figure 2.2 where the noise of the generative model is twice that of the top row. Note that the posterior uncertainty around the nondecreasing curve is visibly smaller than the unconstrained spline. The unconstrained model can fit to the noise of the data instead of the underlying true function. In the bottom row, the mean of the nondecreasing spline almost match the spirit of true function. Finally, we of course see that inappropriate constraints (nonincreasing) force the resulting model to fail entirely to model the underlying phenomena.

Why discount past information? Often, forecasts of future returns use all historical data, equally weighted. However, if the function of interest changes over time, then the more time between a past observation and the future time of interest, the less relevant that observation is.

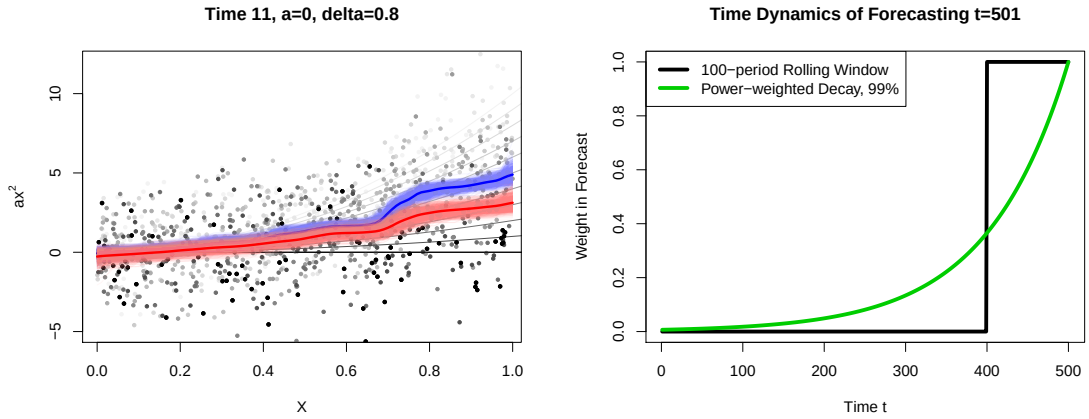
As an example consider the function $f(x, a) = ax^2$ as $a \to 0$. In Figure 2.3, we plot this parabola for $a \in \{10, ..., 1, 0\}$, such that at time 1 the function is $10x^2$, and at time 11 the function is 0. This is a parabola flattening over time, as illustrated by the

Figure 2.2: Fits to simulated data with different amounts of noise



Data generated from a monotonically increasing mean function. Shown are three spline fits to the simulated data: (left) nonincreasing quadratic splines, (middle) quadratic splines without constraints, and (right) nondecreasing quadratic splines. The top row is a low noise environment, and the bottom row is a high noise environment.

Figure 2.3: Visualizations of discounting data over time



(left) Generated data from a parabolic function which collapses to a constant function over 11 time points. White to black points display the simulated data, with darker colors corresponding to points generated more recently. Also displayed are the function fits when treating each time point equitably (blue) versus having decaying weights for increasingly further away time points (red). (right) Example weighting scheme for rolling window weighting (black) versus time-decay weighting (green).

random data points and their mean generative curve are fading from black to white. The pink curves are the draws from a power-weighted-discounted model with $\delta = 0.8$, where at time τ, $\omega_t = 0.8^{\tau - t}$, which implies an [asymptotic] effective sample size of $\frac{1}{1-\delta} = 5$, and the red curve is the posterior mean curve. The light blue lines are the MCMC sample curves from a historic-window model (all past time-periods are equal weighted, so sample size is 11 time points), and the blue curve is their posterior mean. As displayed in Figure 2.3, allowing for time-variation permits the model to better track the current state of a relationship that changes over time, as the red curve is closer to current function (flat) than the blue curve is.

It is important to highlight the differences between a rolling window model and our proposed alternative. First, rolling window is a special case of this model (see McCarthy and Jensen (2016)). Second, completely forgetting past data is not a desirable property. While older data is clearly not as valuable or pertinent as recent data, its value is not zero. Furthermore, if a 10 year rolling window is used, then data from 120 months ago is valued the same as today's, while data from 121 months ago is thrown away, as shown in the figure

below. The arbitrary cutoff between 120 and 121 months does not reflect the true value of information on either side of that threshold. We propose that in the case of time-varying phenomena, the importance of data decays as the data ages, akin to our power-weighting specified above. The exception to this are structural shocks that may occur, but even a 120-month rolling window will take 120 months to fully adapt. If adapting to shocks is the desired property, a structural break model should be used (e.g. Pettenuzzo and Timmermann, 2011).

2.4 Selection Methodology

This section develops an approach for selecting meaningful firm characterstics. The aim is to identify characteristics of a firm that are predictive of its return, and how this set varies in time. This approach builds upon the decision-theoretic selection procedure first proposed in Hahn and Carvalho (2015) and developed for econometric applications in Puelz et al. (2017b,a).

Rewriting the model as a predictive regression. As a first step, we rewrite our model as a predictive regression. Focusing on time t in the cross section, the fully specified model for the vector of n_t firm returns R_t is:

$$\begin{aligned}
\mathrm{R}_t &\sim \mathrm{N}\left(\alpha_t \mathbf{1}_{n_t} + \mathbf{X}_{t-1}\boldsymbol{\beta_t},\ \sigma_t^2 \mathbb{I}_{n_t}\right) \\
\alpha_t &\sim N(0, 10^{10}) \\
\sigma_t^2 &\sim U(0, 10^3) \\
(\gamma_{jkt} | I_{jkt} = 1) &\sim N_+(0, c_k \sigma_t^2) \\
(\gamma_{jkt} | I_{jkt} = 0) &= 0 \\
I_{jkt} &\sim Bn(p_{jkt} = 0.2)
\end{aligned} \tag{2.12}$$

where $\mathbf{X}_{t-1}\boldsymbol{\beta}_t = \mathbf{X}_{t-1}\mathrm{diag}_K(\mathbf{L})^{-1}\mathrm{diag}_K(\mathbf{L})\boldsymbol{\beta}_t = \mathbf{W}_{t-1}\boldsymbol{\gamma}_t$. Note that $\mathrm{diag}_K(\mathbf{L})$ is a block diagonal matrix of size $K(\grave{m} + \acute{m} + 3) \times K(\grave{m} + \acute{m} + 3)$ where each lower triangular block is the projection matrix $\mathbf{L}$. Also, $\mathbf{X}_{t-1}$ is matrix of size $n_t \times K(\grave{m} + \acute{m} + 3)$ and $\boldsymbol{\beta}_t$ is vector

of size $K(\grave{m} + \acute{m} + 3)$. Therefore, each firm is given a row in $\mathbf{X}_{t-1}$, and each $\grave{m} + \acute{m} + 3$ block of $\boldsymbol{\beta}_t$ corresponds to the coefficients on the spline basis for a particular characteristic, k. Incorporating the intercept directly into the characteristic matrix, we can write the generating model compactly as:

$$\mathrm{R}_t \sim \mathrm{N}\left(\mathbb{X}_{t-1}\mathbf{B}_t,\ \sigma_t^2\mathbb{I}_{n_t}\right), \tag{2.13}$$

where $\mathbb{X}_{t-1} = \begin{bmatrix}\mathbf{1}_{n_t} & \mathbf{X}_{t-1}\end{bmatrix}$ and $\mathbf{B}_t = \begin{bmatrix}\alpha_t & \boldsymbol{\beta}_t\end{bmatrix}$.

After rewriting our model more compactly, we delve into the second main contribution of this paper – firm characteristic selection in light of uncertainty. As described in the introduction, there are many firm characteristics available for predicting returns. This leads to a natural question, which small subset of characteristics is *most relevant* for predicting the cross section of firm returns? Further, does this subset vary over time?

Two components: Predictive uncertainty and utility. Suppose we have fit Model (2.12) using standard Monte Carlo methods. We now have access to the posterior distribution over all parameters: $p(\Theta_t \mid \text{past data} = \mathbf{R}_t)$. Also, conditional upon these posterior draws, we can simulate from the predictive distribution, providing draws from the joint distribution of future firm returns $\tilde{\mathrm{R}}_t$ and model parameters Θ_t, written as: $p(\tilde{\mathrm{R}}_t, \Theta_t \mid \text{past data} = \mathbf{R}_t)$. Uncertainty from the predictive is the first input for the selection procedure.

The second component is a rule for comparing models to one another – we call this our utility function. With both predictive uncertainty and a utility function in hand, we can ask and answer the pivotal question: *In light of uncertainty*, how do simpler models with fewer characteristics compare to the model including all characteristics? The decision-theoretic blend of these two components, a Bayesian model and a utility function, will allow us to discern which characteristics are important while taking uncertainty of all forms into account.

Optimizing expected utility and model selection. We formalize this methodology by first deriving our expected utility (loss) function. A natural utility function is the log density of Regression (2.13). Note that Regression (2.13) is not being used here in a statistical capacity for model estimation, but rather as a measure of how well a sparse representation of the linear predictor represents future data. The log density may be written as:

$$\mathcal{L}(\tilde{\mathrm{R}}_t, \mathbf{A}_t) = \frac{1}{2}(\tilde{\mathrm{R}}_t - \mathbb{X}_{t-1}\mathbf{A}_t)^T(\tilde{\mathrm{R}}_t - \mathbb{X}_{t-1}\mathbf{A}_t) \tag{2.14}$$

where $\tilde{\mathrm{R}}_t$ is future return data at time t and $\mathbf{A}_t$ is the "action" to be taken by the data analyst. This action is intended to represent a sparse summary of the regression vector $\mathbf{B}_t$. In order to encourage sparsity in $\mathbf{A}_t$, we include an additional penalty function Φ with parameter λ_t:

$$\mathcal{L}_{\lambda_t}(\tilde{\mathrm{R}}_t, \mathbf{A}_t) = \frac{1}{2}(\tilde{\mathrm{R}}_t - \mathbb{X}_{t-1}\mathbf{A}_t)^T(\tilde{\mathrm{R}}_t - \mathbb{X}_{t-1}\mathbf{A}_t) + \Phi(\lambda_t, \mathbf{A}_t). \tag{2.15}$$

We now integrate the loss function over all uncertainty given by the predictive distribution of asset returns conditioned on observed data: $p(\tilde{\mathrm{R}}_t \mid \mathbf{R}_t) = \int p(\tilde{\mathrm{R}}_t \mid \Theta_t, \mathbf{R}_t)p(\Theta_t \mid \mathbf{R}_t)d\Theta_t$. We do this integration in two steps, first over $\tilde{\mathrm{R}}_t \mid \Theta_t$ and second over Θ_t:

$$\begin{aligned}\mathcal{L}_{\lambda_t}(\mathbf{A}_t) &= \mathbb{E}_{\Theta_t}\mathbb{E}_{\tilde{\mathrm{R}}_t|\Theta_t}\left[\frac{1}{2}(\tilde{\mathrm{R}}_t - \mathbb{X}_{t-1}\mathbf{A}_t)^T(\tilde{\mathrm{R}}_t - \mathbb{X}_{t-1}\mathbf{A}_t) + \Phi(\lambda_t, \mathbf{A}_t)\right] \\ &\propto 2\overline{\mathbf{B}}_t^T\mathbb{X}_{t-1}^T\mathbb{X}_{t-1}\mathbf{A}_t + \mathbf{A}_t^T\mathbb{X}_{t-1}^T\mathbb{X}_{t-1}\mathbf{A}_t + \Phi(\lambda_t, \mathbf{A}_t) + \text{constants}.\end{aligned} \tag{2.16}$$

After integration, we notice that the posterior mean of the coefficients, $\overline{\mathbf{B}}_t$, appears in the first term, while the expectations pass over the second and third terms.

We complete the square and drop constants to obtain the final form of the integrated loss function:

$$\mathcal{L}_{\lambda_t}(\mathbf{A}_t) = \left\|\mathbb{X}_{t-1}\mathbf{A}_t - \mathbb{X}_{t-1}\overline{\mathbf{B}}_t\right\|_2^2 + \Phi(\lambda_t, \mathbf{A}_t) \tag{2.17}$$

For a fixed time t, Loss (2.17) has the same form as the one derived for linear regression models in Hahn and Carvalho (2015). The third and final step is to choose a penalty function Φ and optimize the loss function for a range of λ_t for each time t.

For this paper, we choose $\Phi(\lambda_t, \mathbf{A}_t) = \lambda_t \sum_{k=1}^{K} \left\| \mathbf{A}_t^k \right\|_2^2$ where $\mathbf{A}_t^k$ is the k^{th} $\grave{m} + \acute{m} + 3$ block of the vector $\mathbf{A}_t$ after neglecting the intercept. The group lasso algorithm of Yuan and Lin (2006) is then used to minimize the integrated loss. This provides a way to jointly penalize groups of covariates. In the context of our financial application, this "group penalization" permits the selection of firm characteristics by grouping the coefficients of a single quadratic spline together in the penalty.

In order to see this, recall the structure of the sparse action $\mathbf{A}_t$. It is a $K(\grave{m}+\acute{m}+3)+1$ length vector where the k^{th} $\grave{m} + \acute{m} + 3$ block (excluding the intercept) corresponds to the spline basis for firm characterstic k. By using the approach outlined in Yuan and Lin (2006), we group together the spline bases for each characteristic. Then, Loss (2.17) is minimized for varying penalty parameter choices, such that we can look at a range of quadratic spline models built from one characteristic up to the 36 characteristics available.

Posterior summary plots. These sparse models are optimal under our choice of utility and fixed level of regularization given by the penalty parameter, and we can compare them in light of the statistical uncertainty from the Bayesian model. Denoting the collection of sparse optimal models $\{\mathbf{A}_{\lambda_t}^*\}$, we study the distribution of the *difference in loss* of a reduced model and the full model:

$$\Delta_{\lambda_t} = \mathcal{L}(\tilde{\mathrm{R}}_t, \mathbf{A}_{\lambda_t}^*) - \mathcal{L}_0(\tilde{\mathrm{R}}_t, \mathbf{A}_0^*) \tag{2.18}$$

where $\mathcal{L}$ is as defined in Equation (2.14). Note that, as $\mathcal{L}$ is a random variable, so is Δ. Crucially, this metric incorporates statistical uncertainty through the predictive and optimality through consideration of the set $\{\mathbf{A}_{\lambda_t}^*\}$.

An important feature of this approach is the ability to identify important return predictors and how this set may vary *over time*. The time variation and connection across time periods is driven by the power-weighted density approach and embedded in the posterior (recall that the rolling-window model is a special case of the power-weighted density approach).

Therefore, although the minimization of the integrated loss is performed myopically at each point in time, the variation of optimal sparse models across time may be studied.

2.5 Case Study

Our case study focuses on a rich data set from Freyberger et al. (2019). It is a monthly panel where we observe a cross section of firms, their excess return as well as 36 lagged characteristics of each. The full dataset spans 623 months of returns, July 1962 through May 2014 and includes 1,404,048 observations. We train our models on the first 12 years, and then test and update the models on the remaining data. Thus, the results shown cover 1974-2014. The posterior distributions are updated annually.

The characteristics are listed in Table D.1 in Appendix D as well as the direction of monotonicity we impose for each. We will examine three different sets of splines: those with no constraints (nonmonotonic), some constraints, and many constraints. The model with some monotonic constraints applies the rather established evidence from the financial literature (Fama and French, 2016) and constrains size, book-to-market, profitability, investment, momentum (Jegadeesh and Titman, 1993, 2001), and intermediate momentum to be monotonic. For the "fully monotonic" model with many constraints, we impose monotonicity constraints on every variable whose constraint had reasonable support in the literature; thus 24 of the 36 variables are constrained. The supporting papers are also listed in Table D.1.

As benchmarks, we fit a 120-month rolling window OLS and a 120-month rolling window random forest – a nonparametric ensemble learning model with competitive predictive ability across many applications (Breiman, 2001). We also fit 12 specifications of our additive quadratic splines model across the 3 different sets of monotonic constraints mentioned above and across 4 different specifications of discounting: no discounting (historic window), 120-month rolling window, and power-weighted discounting with $\delta = 0.990$ as this has an

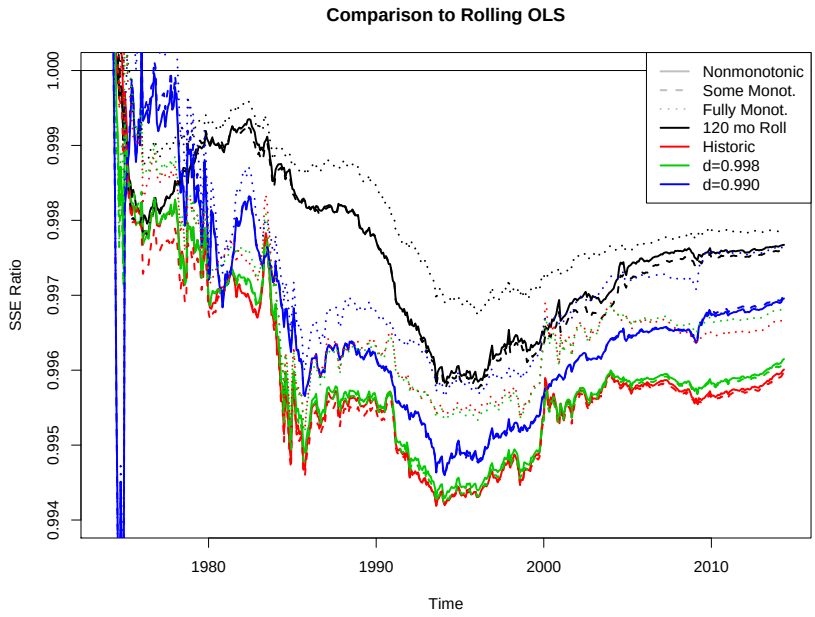
effective sample size of 100 months, or $\delta = 0.998$ is 500 months, and thus is only slight discounting.

Three sections follow. First, we detail the modeling results. Second, we present characteristic selection results (conditional upon the models). Third, we discuss the economic ramifications of the proposed approach.

2.5.1 Modeling Results

The impact of model specification. In this section, we display the forecasting ability of the 12 model specifications considered. Since our selection process requires a posterior distribution as an input, this analysis helps in identifying the model (equivalently, the monotonicity and time-variation specifications) among our 12 that possesses the best predictive ability. The curves in Figure 2.4 are the aggregate sum of squared errors (SSE) up to time t for a model, as a percentage of the aggregate SSE of the OLS model. A pattern exists across both time-discounting amount and number of monotonic constraints. We see that the dotted lines representing the models with many monotonic constraints are always higher than the dashed and solid lines. Hence, a large number of monotonic constraints may not be of benefit to forecast errors, at least compared to a more moderate number of constraints. We see that using some monotonic constraints, as represented by the dashed lines, has slightly better point forecasts than the solid curves that represent the nonmonotonic models. Figure 2.4 also shows the effect of different time windows and time-discounting. Essentially, point forecast success prefers little discounting of past information: using all the data is preferable to slight discounting, which is preferable to heavy discounting and rolling windows. Note that the random forest model is not shown, as its SSE curve is off the top end of the chart. Random forest SSE is about 1.5% greater than OLS (1.015 on the shown y-axis, hence not plotted as it's largely not visible). These results are echoed in the full-term RMSE's given in Table 2.1.

Figure 2.4: Aggregate squared error ratio over time



Aggregate squared error ratio, over time. The sum of squared forecast errors for a given model, divided by the sum of squared forecast errors of the OLS model.

Table 2.1: Root mean squared prediction errors

	Rolling	Historic	$\delta = 0.998$	$\delta = 0.990$
Random Forest	0.764			
Nonmonotonic	-0.116	-0.200	-0.193	-0.152
Monotonic - Some	-0.120	**-0.202**	-0.195	-0.151
Monotonic - Many	-0.107	-0.165	-0.158	-0.117

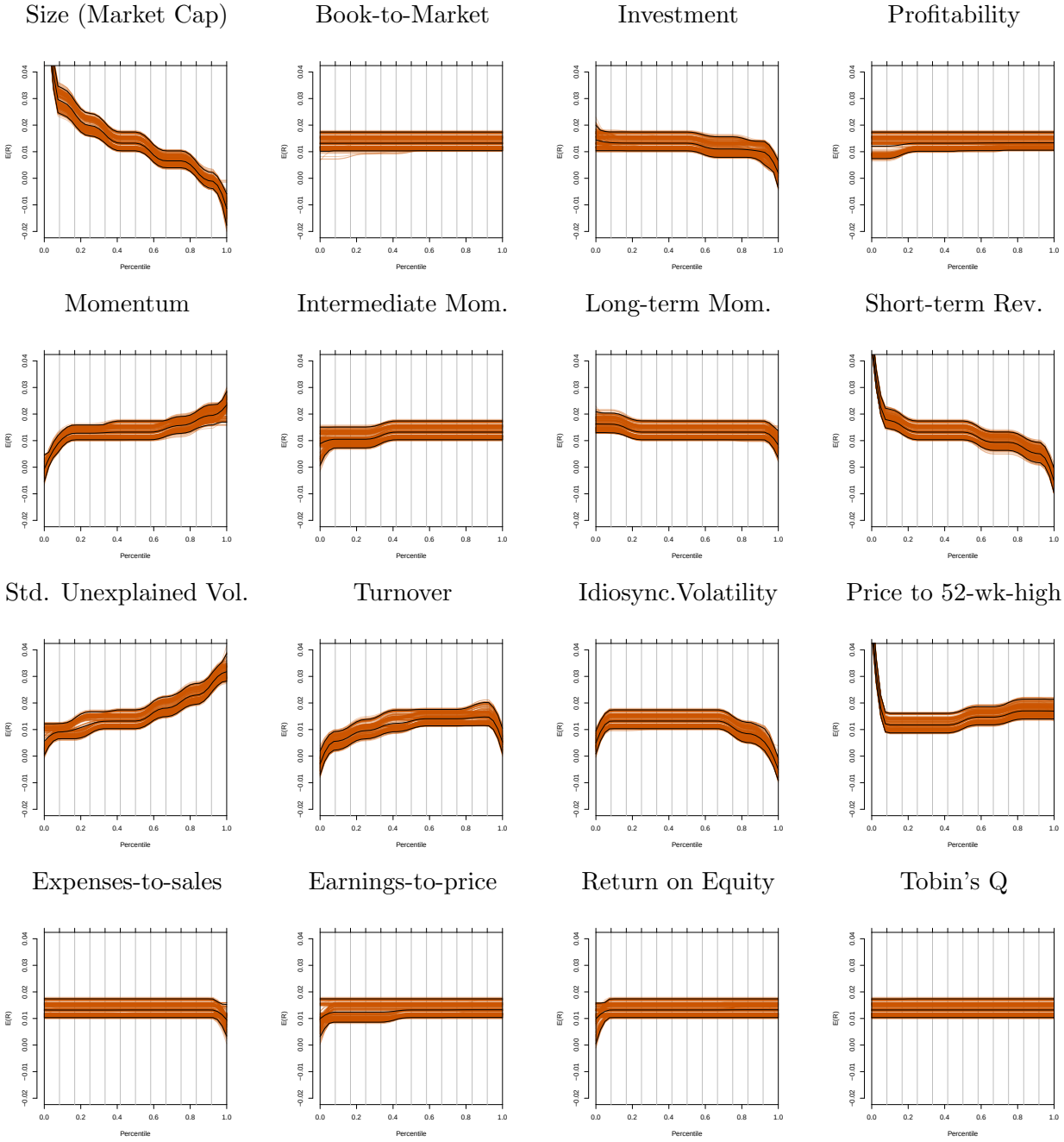
Root mean squared prediction errors over the 1974-2014 period, as percentage change relative to OLS on 120 month rolling window: $100\left(\frac{RMSE_{model}}{RMSE_{OLS}} - 1\right)$. Note that OLS RMSE is 0.172.

The significant underperformance of the random forest model is worth additional comment. The bias-variance tradeoff and low signal-to-noise environment in finance data are key concepts affecting these results. The highly flexible nonlinear model given by the random forest is overwhelmed with noise, and its resulting performance is poor. This underscores a need for structured models in these applications beyond their ease of interpretability. Bias induced by our structured monotonic spline model leads to an outperformance of popular machine learning methods where both structure and interpretability are minimal. In the following sections, we explore the return-characteristic relationships that our fitted models provide.

The return-characteristic relationship. Figure 2.5 shows the partial effects of each of these selected 25 characteristics, from a historic window model. Specifcally, we use the subset of monotonic constraints as it has the best fit in terms of forecast error; see Table 2.1. Each individual pane shows the partial effect of a characteristic assuming the other 35 characteristics are held at their medians. The first thing to note are the strong effects of size, momentum, short-term reversal, standard unexplained volume, and price to 52-week-high. We also see that there are some nonmonotonic effects: turnover, idiosyncratic volatility, and price to 52-week-high. This shows main reason why excessive monotonic constraints can hurt the model: some relationships are not monotonic. Also, we see some effects that are almost zero, such as book-to-market, which is a staple in empirical finance work.

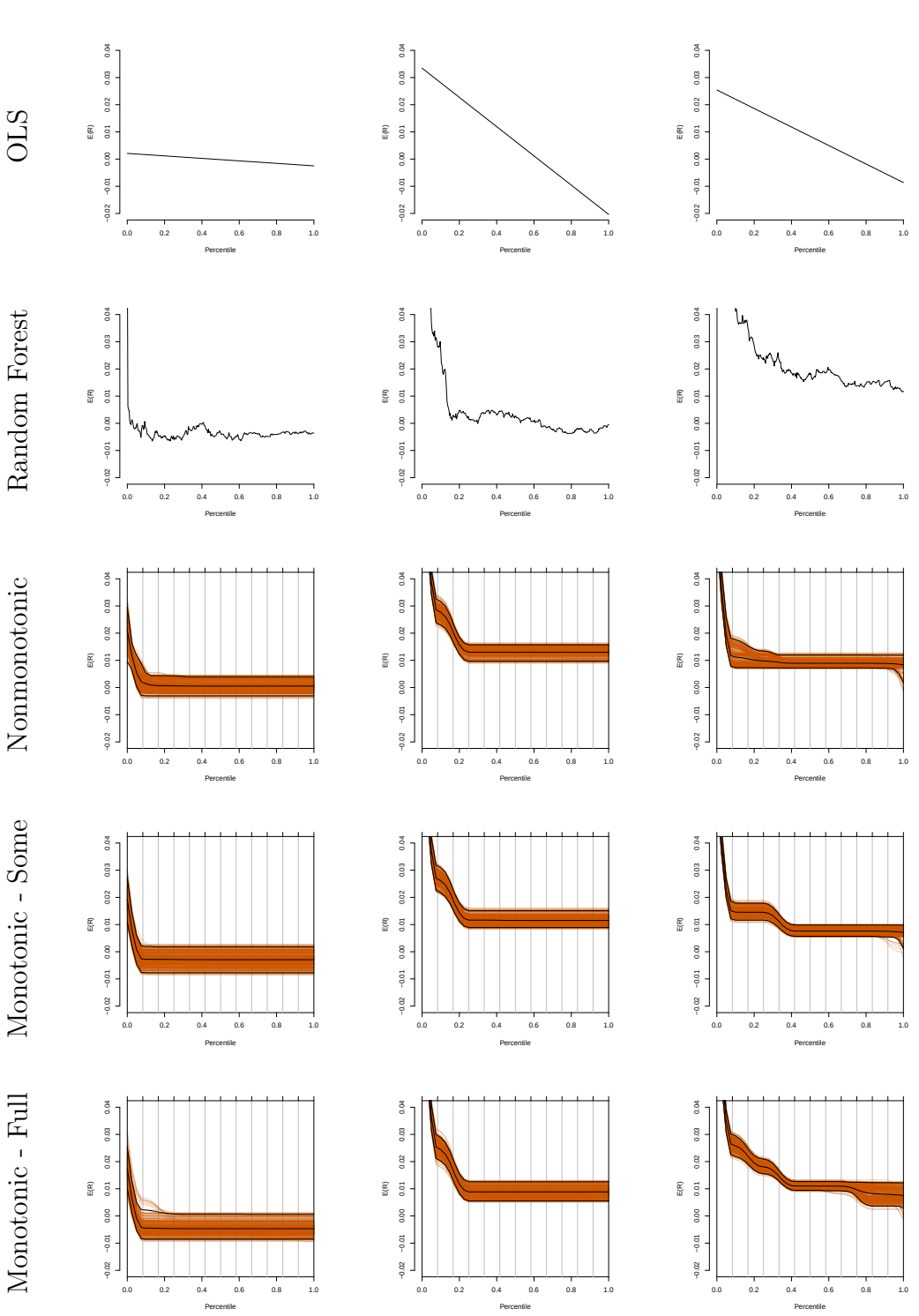
The return-characteristic relationship, over time. We next look at partial effects given at different points in time for the different models, in Figure 2.6. This figure shows the partial effect of firm size on returns when holding all other variables at the median, estimated by multiple linear regression (OLS), a Random Forest, and our different spline models. The effects and their uncertainty are given for January of 1974, 1994, and 2014, each using a 120 month window of training data (rolling window models). Generally, we see the size effect growing stronger over time: the smallest firms see much larger average

Figure 2.5: Effects of characteristics on returns



Effects of characteristics on returns over the historic window of 1974-2014 period (each observation equally-weighted over time), ordered according to the order of inclusion. Here, only six variables are constrained to be monotonic: size, book-to-market, investment, and profitability on the first row, as well as momentum and intermediate momentum on the second row. The remainder of the second row is composed of other functions of past returns. The third is composed of various pronounced effects, while the fourth row contains characteristics with much smaller or no effects in the full posterior. The three black curves are the posterior mean and the 95% credible bands. The transparent orange curves are each of the MCMC draws, such that darker orange areas reflect greater posterior density. The vertical gray lines show where the knots are placed. The horizontal axes are the percentiles of the characteristic. The vertical axes are the expected returns.

Figure 2.6: Comparisons of effect of firm size at different points in time

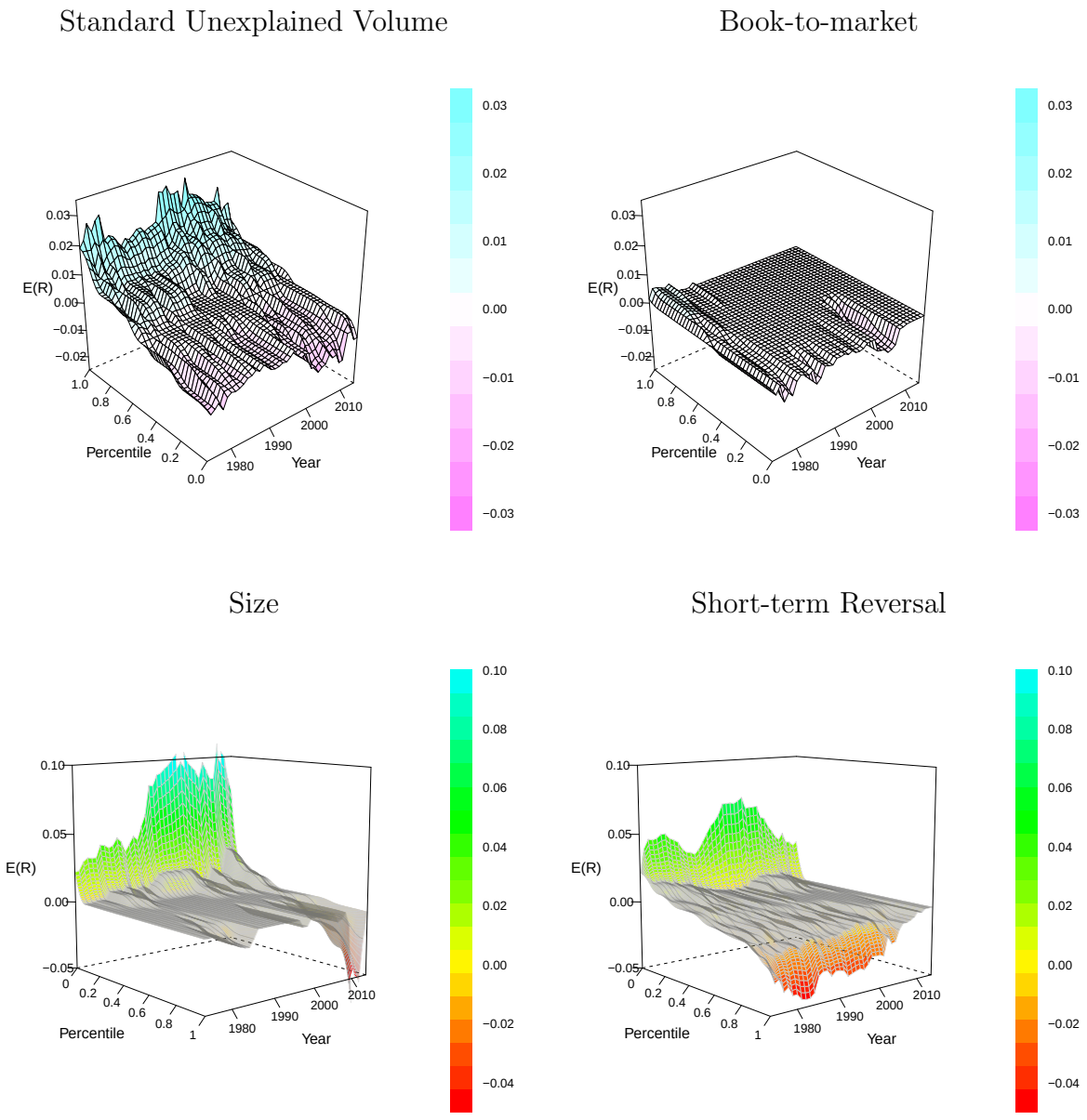


Comparisons of functions of size (market cap) at different points in time. Assumes all other variables are held at their medians (0.5). Each uses a 120 month rolling window. The horizontal axes are the percentiles of size. The vertical axes are the expected returns.

returns than all other firms. This effect is blurred by standard regression's assumption of a strictly linear relationship. Random forests pick up this small-firm phenomenon, but the resulting curve is noisy and wiggly. Our splines with shrinkage at the knots smoothes over this noise and avoids overfitting the wiggles. Furthermore, this figure visually demonstrates two sources of uncertainty and variance reduction: more data and more structure (the bias part of the bias/variance tradeoff). There are more firms in 2014 than in 1974, and thus the 95% credible or confidence bands decrease as we move from the left panes to the right panes. The further reduction in variance we see as we progressively move from the middle right panes to the bottom right panes comes from adding monotonic constraints. This is especially seen in the tighter probability bands in the 2014 panes: as more monotonicity constraints are added to the model in general, the estimate of the size effect becomes more certain, even though the splines for size never appear nonmonotonic.

Figure 2.7 shows the annual progression of the splines. There are two different color schemes to denote two different sets of axes. The first panel illustrates a common pattern: little change over time. The effect of standard unexplained volume is fairly consistent over time, with slight fluctuations. Next, while frequently included in our models and most other papers' asset pricing models, book-to-market does not have a very strong effect, though this could change with different control variables. We see some value premium (high expected returns of high book-to-market firms) in the late 1970's, and low returns of growth firms (low percentiles of book-to-market) in the 2000's, which, as this is a 120-month rolling window, likely reflects the burst in the dot-com bubble. The positive returns seen by the smallest firms (size) increase halfway through the period, while there is little effect on the large firms until the Great Recession, as seen in the red dip near the end of last decade. While the effect of short-term reversal (firms performing well last month tend to underperform this month, and vice versa) on the low percentile/worst firms' positive returns are fairly stable over time, the negative effect on returns of last month's winners depletes over time.

Figure 2.7: Effects over important characteristics over time



Splines of most included effects, throughout the test period. For the given characteristic, the splines for January in each year are placed in order, yielding a response surface of the rank-transformed characteristic and time versus monthly expected returns. The blue/pink color scheme has an increasing percentile axis. Colors are assigned to buckets of expected returns 50 basis points wide, such that regions of expected returns between -25 and +25 basis points are white. These basis point changes are with respect to a firm with the median value of the characteristic in a given year. Hence, the white areas of the plot reflect percentiles of firms that do not vary significantly from the median firm. The red/green/cyan color scheme flips the percentile axis so the curve is viewable and zooms out along the E(R) axis essentially doubling the limits and halving the granularity of the color spectrum.

2.5.2 Selection Results

Which characteristics are important? We first look at which characteristics are important over the whole time period, 1974-2014. To do this, we use the historic window model where each observation over the 41-year period receives equal weight. Also, the small set of monotonic constraints is used, as it has the best fit in terms of forecast error; see Table 2.1. Then, posterior summarization is performed as detailed in Section 2.4, except that we look at the whole period as a single time step, thus $\mathbb{X}_{t-1}$ contains data from over the whole time period. We do this by looking at the distribution of the difference in loss of a reduced model and the full model defined in Expression (2.18).

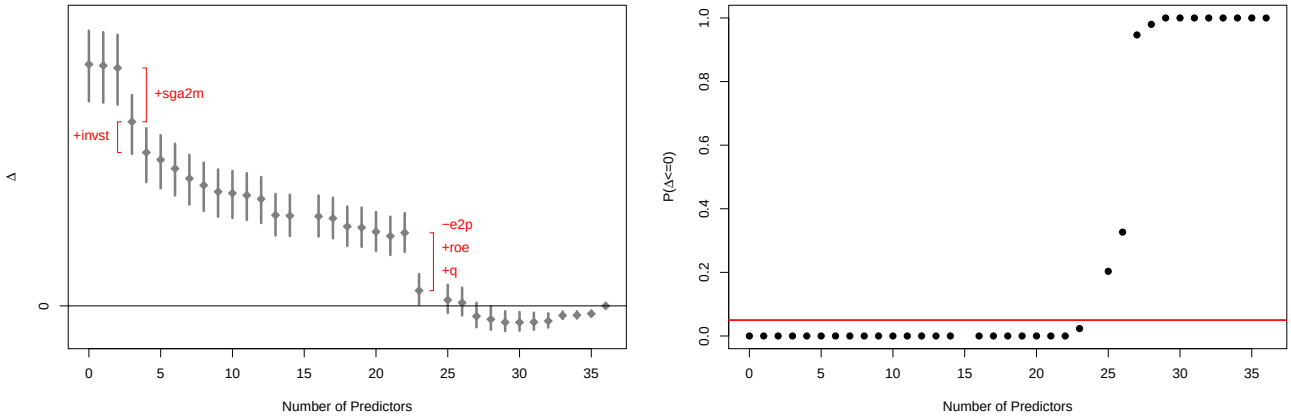
A convenient feature of the selection methodology is the ability to undertake a full sample analysis such as this. Beyond the loss function, the remaining components for the method are the predictive distribution calculated at the end of our sample, and a subsample of our data to build $\mathbb{X}_{t-1}$.[1] Then, the expected loss is computed, optimized, and the optimal sparse models are compared in light of predictive uncertainty. This *separation* of inference from characteristic selection is a helpful tool for exploratory analysis within our case study.

In Figure 2.8, we show a difference in loss metric Δ_{λ_t}, in the left panel, and the probability that a sparse model has less loss than the full model, $P(\Delta_{\lambda_t} < 0)$, in the right panel, for a sequence of models with varying numbers of included characteristics. One can think of λ_t as indexing models of varying sizes. The models in this sequence are minimum loss models for each number of included characteristics.[2] Figure 2.8 shows that using 27 or more characteristics has a very high probability of having the same or smaller loss than the full model. The left panel of Figure 2.8 also shows that the timely inclusion of expenses-to-

[1] We summarize the posterior with respect on the a random month from each of the 41 years in the test set, as using all firm-year observations in $\mathbb{X}_{t-1}$ from Equation (2.13) to summarize the posterior currently does not work on a 16GB RAM machine.

[2] Equation (2.17) is optimized for hundreds of values of λ_t. Of all the models with p covariates selected into the model, the chosen model has the minimum loss among all models with p covariates.

Figure 2.8: Posterior summarization using difference in loss



Posterior summary plots over the full test period. The left panel shows the distribution of the difference in loss for models of differing numbers of characteristics relative to the full model. In red are shown the variables that, when added (+) or removed (−) from the model, cause significant changes in the loss distribution. The right panel shows the probabilities of having the same or better loss than the fully dense model of all 36 predictors. The red threshold in the right panel is 0.05, and the model immediately above the threshold is selected.

sales, investment, return on equity, and Tobin Q's all lead to significant movements (towards zero) in the distribution of Δ_{λ_t}.

The red line in the right panel shows our threshold of 0.05, meaning models above the threshold have at least a 5% chance of having equal or better loss than the full/dense model. We select the sparsest model over the threshold, which has 25 characteristics. These are given in Table 2.2 in order of inclusion. Table 2.2 shows us that while the variables from Fama and French (2016) are present (i.e. investment, book-to-market, size, and profitability), they do not come first – standard unexplained volume and short-term reversal are the first characteristics to enter the sparsest models. Both of these have large effects over the sample, as we saw in Figure 2.5.

When are characteristics important? To answer our second question, we implement the same posterior summarization as mentioned previously, but now annually. Using the

Table 2.2: Variables selected via posterior summarization

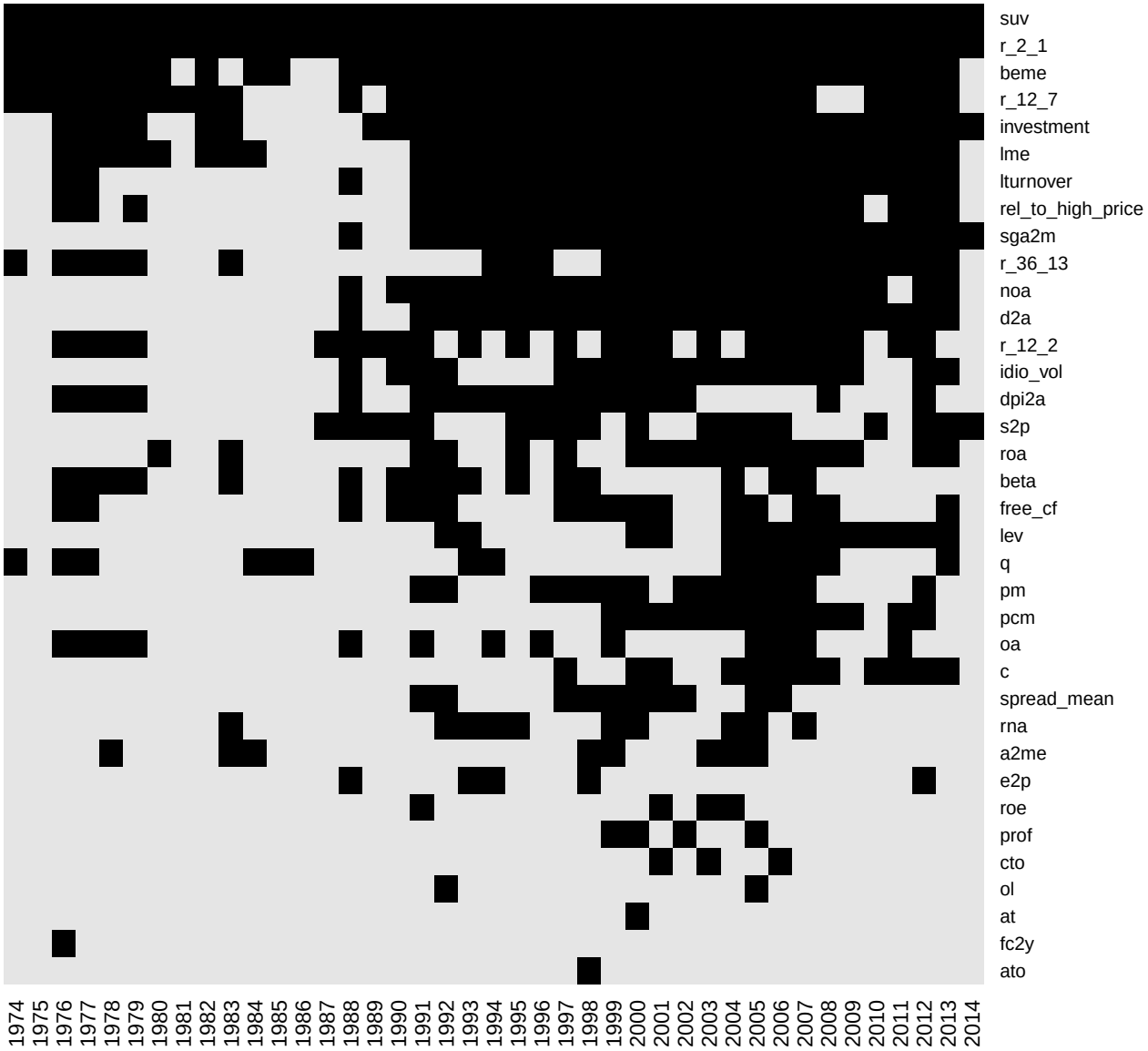
1. Standard unexplained volume	14. Free cashflow
2. Short-term reversal	15. cash-to-assets (tie)
3. Expenses-to-sales	15. price-to-cost margin (tie)
4. Investment	17. Price to 52-week-high
5. Book-to-market	18. Return on assets
6. Momentum	19. Idiosyncratic volatility
7. Intermediate momentum	20. Profit margin (tie)
8. Size (Market cap)	20. Sales-to-price (tie)
9. Depreciation-to-assets	22. Tobin's Q (tie)
10. Long-term momentum	22. Return on equity (tie)
11. Net-operating assets	24. Change in PP&E and inventory (tie)
12. Turnover	24. Profitability (tie)
13. Leverage	

Selected variables variable for a threshold of 0.05 on $P(\Delta_{\lambda_t} < 0)$, ordered by order of inclusion in the model. These are variables included from the model that does no worse than the full model with 5% probability. (tie) denotes variables that come in to the model at the same time. Note that in Figure 2.8, there is no optimal model with 15 or 24 predictors, thus why we see ties at 15th and 24th above. The other ties are instances of two variables coming into the model as another leaves. However, variables that leave the model and do not come back in for the ideal set of 25 are not included in the overall ranking.

same 0.05 threshold and employing the posterior form the partially monotonic, lightly time-discounted model ($\delta = 0.998$), we plot the selected covariates in Figure 2.9. Black cells indicate selected characteristics and light grey cells represent excluded characteristics. The figure visibly has two separate periods, and the transition happens over, or somewhere in, the 1987-1991 range. Here, we define the break to be between 1990 and 1991, as there are never fewer than 14 characteristics selected from 1991-2013. Thus, from the beginning of the evaluated data until 1990, a small number of characteristics were selected annually, between 3 and 16. During the second period, 1991 to present, a larger number of characteristics were selected annually, between 14 and 28. The exception is 2014, which dips down to pre-1990 levels with 5 selected characteristics. This paper does not put us in a position to make a causal statement as to why these changes happen, but we will comment on what happens.

In regards to specific characteristics, we first note that standard unexplained volume and short-term reversal are the only variables selected every year. During the first period (1974-1990), book-to-market and intermediate momentum are also selected in more than

Figure 2.9: Variable selection over time



Using the same 0.05 threshold, we find the sparsest model with at least 0.05 probability of having no more loss than the fully dense model, for every year (January). Variables on the vertical axis are ordering according to the frequency of their appearance. Black cells indicate selection, while light grey cells indicate exclusion from the sparse model.

half of those 17 years, with size, investment, and momentum coming close to that mark. In the main piece of the second period (1991-2013), in addition to the two mainstays, the variables selected every year are book-to-market, depreciation-to-assets, investment, size, turnover, and expenses-to-sales. 14 other characteristics were selected during more than half of these years. In 2014, only five characteristics were selected: standard unexplained volume, short-term reversal, investment, expenses-to-sales, and sales-to-price. In reference to the literature (Jegadeesh and Titman, 1993, 2001; Fama and French, 2016) and our smaller set of monotonic constraints, there appears to be support for size, book-to-market, investment, momentum, and intermediate momentum, but not for profitability. Perhaps there's a characteristic, or combination of characteristics, that better fits the data and the notion of "profitability" than does the characteristic that we use.

2.5.3 Are there Economic Gains?

This final section considers the economic impact of the proposed methodology. The challenge with looking too closely at point forecasts is that slight differences may not matter much: the signal to noise ratio is small. An alternative approach for model comparison is to compute portfolio metrics. Thus, we consider the annualized Sharpe ratios for equal- and value-weighted decile portfolios in Table 2.3. We take the forecasts from a single model and buy the top decile of stocks by shorting the bottom decile. The stocks in these purchases are either equally weighted or weighted by the future expected return. The annualized Sharpe ratios are from the monthly returns from these long/short decile portfolios. Here, it's quite clear that using many monotonic constraints yields the highest Sharpe ratios. Hence, while they not provide the most absolutely accurate model in terms of SSE, fully monotonic models more correctly indicate the future success of firms. This implies they are more accurate at estimating the tails of the expected returns distribution, i.e. the top and bottom decile.

Table 2.3: Annualized Sharpe Ratios

Equal-weighted Portfolios

	Rolling	Historic	$\delta = 0.998$	$\delta = 0.990$
OLS	2.31			
Random Forest	2.28			
Nonmonotonic	3.02	3.09	3.07	3.01
Monotonic - Some	2.96	3.10	3.17	3.06
Monotonic - Many	**3.22**	3.21	3.19	3.08

Value-weighted Portfolios

	Rolling	Historic	$\delta = 0.998$	$\delta = 0.990$
OLS	2.30			
Random Forest	2.13			
Nonmonotonic	2.88	2.98	2.93	2.90
Monotonic - Some	2.83	2.95	2.98	2.91
Monotonic - Many	3.04	**3.09**	3.06	3.03

Annualized Sharpe Ratios from long/short decile portfolios. "Rolling" is the 120-month rolling training sample. "Historic" uses all past data in the training sample. The remaining columns reflect power-weighted likelihood approach, with the listed discount factor δ. Note that the historic window equivalently has $\delta = 1.0$. The topic panel reflects equal weights, while the bottom panel shows results from value-weighted portfolios. These weights are made proportional to the forecasted expected returns of the top decile and bottom decile separately.

2.6 Conclusion

The intersection of flexible modeling in Bayesian statistics and characteristic selection in finance is the focus area of this paper. We develop a statistical method for modeling returns based on the joint distribution of characteristics as well as provide a way to identify significant ones in light of statistical uncertainty. Our case study concludes that thoughtful model construction is important when dealing with finance data. Comparisons of modeling approaches cautions against use of highly flexible machine learning methods. Our conclusions suggest that model structure (through additivity and monotonicity) provides the dual benefit of interpretability and enhanced predictive ability.

Specifically, there are three important contributions made by our model in this paper. First, our flexible and interpretable model is Bayesian, and thus accounts for the different sources of uncertainty. Second, the model supplements the flexibility of quadratic splines with theoretically-supported monotonic constraints, being one of the least imposing forms of structure. Third, we modify Shively et al. (2009)'s monotonic splines to be time-dependent in order to model the nonlinear yet possibly-dynamic relationships of returns and characteristics. We carefully investigate time variation our model using the methods of McCarthy and Jensen (2016) to discount past data. We find strong evidence for monotonicity even after conditioning on many other available characteristics. This conclusion is supported statistically and economically by an analysis across 12 model specifications.

The fourth contribution, and the second half of this paper, is the development of a utility-based selection procedure for our model. Using this new approach, we are able to uncover the practical significance of characteristics and how these effects vary in time. We find about two dozen firm characteristics that have been important over the last four decades. However, we note that the timing of the importance of each characteristic varies, and two, that the magnitude of each characteristic's effect ranges from negligible to large, and this too can vary over time. With these methods, we find that characteristics with the

largest effects on expected returns are size, short-term reversal, and standard unexplained volume. We find that, while the specifics of these effects change over time, their importance does not diminish. We also find that book-to-market, investment, and momentum are also important over all time, although their effect sizes in the full posterior are not nearly as large as the former three.

Chapter 3

Predicting Counterfactuals and Measuring Heterogeneous Effects of Continuously-distributed Treatments

This chapter reflects my work in progress with Carlos M. Carvalho and Jared S. Murray. We present a novel method for measuring the heterogenoues causal effects of different amounts/levels of a treatment. This method uses Bayesian Additive Regression Trees (BART) priors as they are highly flexible, capable of fitting nonlinearities, interactions, and discontinuities. This high degree of flexibility is tempered by the chosen model structure and the regularization of the priors.

3.1 Introduction: The Causal Problem

The central question in many research fields is if a change in an independent variable, X, causes a change in a dependent variable, Y. This is a question of not just correlation, but causation. We can look at causality through the potential outcomes framework of Imbens and Rubin (2015). Let $Y_i(Z_i)$ represent the outcome for subject i when she receives treatment Z_i. Thus, $Y_i(1)$ is the subject's response when treated and $Y_i(0)$ is the response when not receiving treatment. Note that capitalized Y here is a random variable (as opposed to realized values in lowercase) and it follows that the typical quantity of interest is an expectation, the average treatment effect $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$. For example, a foundational question in empirical corporate finance research is whether or not going into debt ("leverage") increases the profitability of a firm. However, we only get to see the results a single choice per firm, and thus we need to predict (or forecast) the result of the unobserved choice, which is termed

the counterfactual.

Yet, there are two underlying issues. First, regression of Y on Z doesn't work as there can be confounding/control variables, X. Naturally, this is not an issue if a well-designed experiment was conducted such that treatment Z was chosen at random for all subjects. However, it is reasonable that different types of subjects react differently to a treatment, termed heterogeneous treatment effects, and this is the second issue. It is simple to surmise that the profitability of different kinds of firms in different industries may react differently to changes in leverage. Hence, measuring heterogeneous treatment effects can help us understand this phenomena, and others, in a greater light.

With these two issues in hand, we want to carefully handle other variables, which we'll call X. In this chapter we assume all potential confounding variables are included in the $n \times p$ matrix X. Hence under the general assumptions (*strong ignorability* and *overlap*) detailed in Hahn et al. (2018b, Equations 1-3), we get

$$\tau(x_i) := \mathbb{E}(Y_i|\boldsymbol{x}_i, Z_i = 1) - \mathbb{E}(Y_i|\boldsymbol{x}_i, Z_i = 0). \tag{3.1}$$

We see that τ is a function of $\boldsymbol{x}_i \in \mathbb{R}^p$, implying that the treatment effect can change for different values of the covariates.

3.2 The Case for Machine Learning and Regularization

Our main interest is the functional form of $\tau(x)$. In doing so, our work will differ from the papers we reference in one of three ways. First, we allow $\tau(x)$ to flexibly model a variety of complexities. Second, we bridle said flexibility through regularization to avoid confounding and overfit. Third, we will let Z have multiple values.

Hill (2011) notes that, with appropriate assumptions, estimation of binary treatment effects is simply modeling the needed response surfaces. By estimating a surface (e.g. $\mu(x)$) for each of the treatment and controls groups, the heterogeneous treatment effect for a given

$\boldsymbol{x}$ is the difference between the two surfaces at $\boldsymbol{x}$, e.g. $\mu_{treat}(\boldsymbol{x}) - \mu_{control}(\boldsymbol{x})$. These surfaces may contain complexities such as nonlinearities, discontinuities, and interactions (hereafter referred to simply as "complexities"). Linear regression, while widely-accepted, is incapable of estimating these complexities without knowing to look for them *a priori*. On the other hand, machine learning methods excel at fitting these complicated response surfaces, with fewer *a priori* assumptions, and thus can assist with this task. Hill (2011) addresses this challenge through a causal-oriented Bayesian version of Breiman (2001)'s random forest.

The challenge with highly-flexible modeling techniques is the possibility of overfit. This can be combated with regularization, the most famous of which in the simple regression context is the LASSO (Tibshirani, 1996), which shrinks the values of regression coefficients to improve out-of-sample performance. Yet, Hahn et al. (2018a) and Hahn et al. (2018b) explore this in detail, and find that regularizing naively can resort in distorted estimations of treatment effects, both in the linear and machine learning cases. Hence, regularization must be performed carefully, and the work in this chapter follows the guidance in Hahn et al. (2018b).

These papers all focus on a binary treatment, $Z_i \in \{0, 1\}$, where subjects either received a treatment or not. Yet, we are often interested in a dose response: how subjects react to a different amount of a treatment, such that $Z_i \in \mathbb{R}$. These treatments could either be milligrams of a drug, or percent of assets levered to bring in additional funds at a firm. Thus, this chapter extends the work of Hahn et al. (2018b) to cases where the treatment can take on many different values.

3.3 Methodology

The problem at hand is to model the relationship between an observed scalar response variable, $y_i \in \mathbb{R}$, and an observed scalar treatment amount $z_i \in \mathbb{R}$, in the presence of observed, potentially-confounding control variables in vector $\boldsymbol{x}_i \in \mathbb{R}^p$, where $i = 1, ..., n$

subjects. In other words, we need to estimate a potentially-complex function f

$$y_i = f(\boldsymbol{x}_i, z_i) + \epsilon_i \tag{3.2}$$

for subject i and some error ϵ_i. As in Hahn et al. (2018b), we assume f consists of a linear relationship between y_i and z_i, given covariates $\boldsymbol{x}$

$$y_i = \mu(\boldsymbol{x}_i) + \tau(\boldsymbol{x}_i)z_i + \epsilon_i \tag{3.3}$$

where the coefficients μ and τ are functions that map $R^p \to R$, and we make no assumptions about this surface. Thus, the estimation of μ and τ must be flexible across the aforementioned complexities. We submit that BART priors are excellent for this.

3.3.1 BART: Bayesian Additive Regression Trees

Bayesian Additive Regression Trees (BART) is a compilation of weak learners, specifically trees. BART is Bayesian in that there are priors on the tree depth and the leaf values. BART uses a sum of trees, namely

$$y_i = \sum_{j=1}^{m} g(\boldsymbol{x}_i; T_j, M_j) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2). \tag{3.4}$$

The function g maps $\boldsymbol{x}_i$ to a value in set M_j according to partitions of R^p that are defined by tree T_j.

The genius of BART comes through its prior on the tree size. Chipman et al. (2010) set the probability that a node at depth d is nonterminal is

$$\alpha(1+d)^{-\beta}, \qquad \alpha \in (0,1),\ \beta \in [0,\infty). \tag{3.5}$$

With the default values of $\alpha = 0.95$, $\beta = 2$, this prior encourages small trees, but allows further splitting if the data overwhelmingly requests a split. In contrast with random forest methods (Breiman, 2001) which typically have a stopping criteria of a certain leaf size, the trees produced by BART use the data, regularized by the prior, to build the tree shapes instead of a heuristic.

3.3.2 BCF: Bayesian Causal Forests

Hahn et al. (2018b) call their method Bayesian Causal Forests, or BCF. We assume their same priors on μ and τ. Specifically, the BART prior on μ has the default tree priors from Chipman et al. (2010): 200 trees, $\alpha = 0.95$, and $\beta = 2$. The priors on the leaf nodes are, conditional on the tree, normal with mean 0. The standard deviation of this normal distribution has a half-Cauchy prior distribution with a prior median equal to twice the marginal standard deviation of Y.

The BART prior on τ contains stronger regularization, with 50 trees, $\alpha = 0.25$, and $\beta = 3$. This shrinks strongly toward homogeneous effects, such that the data can specify which heterogeneous effects are significant. The priors on the leaf nodes are again, conditional on the tree, normal with mean 0. However, the standard deviation now follows a half normal prior, with prior median equal to the marginal standard deviation of Y.

Using the BART priors means that our model is estimated with Markov chain Monte Carlo methods (MCMC). These prior assumptions are detailed mathematically in Appendix F, along with some key posterior derivations.[1]

3.3.3 Note on Linearity

The linear relationship between z_i and y_i (given $\boldsymbol{x}_i$) in Equation (3.3) will not be the exact case in most data generating processes. As such, when we estimate $\mu(\boldsymbol{x}_i)$ and $\tau(\boldsymbol{x}_i)$ in our model, we are fitting linear models to a region of the support of $\boldsymbol{x}_i$. In the case that, y_i is linear in z_i given $\boldsymbol{x}_i$, we will recover the data generating process. In cases where its not linear but only monotonic, we will recover something close, though the sign of the relationship will be correct, whether positive or negative. We look at these cases in the following simulation studies.

[1] Due to these current calculations of the MCMC algorithm, z_i cannot be 0 for any i as it appears in some denominators.

3.4 Simulation Studies

3.4.1 Simple Generative Model

To demonstrate the how this model works, we first estimate a simplified version of the generative model with no intercept μ. Note that μ is not estimated in the MCMC sampling either, hence

$$y_i = \tau(x_i)z_i + \epsilon_i. \tag{3.6}$$

Here we take the treatment effect (slope) to be a simple step function:

$$\tau(x) = \begin{cases} -1 & x < 0.5 \\ 1 & x \geq 0.5. \end{cases} \tag{3.7}$$

For $i = 1, ..., n = 200$, we generate x_i and z_i from a Uniform(0,1) distribution and ϵ_i from a Normal$(0, \sigma^2)$ distribution with $\sigma = 0.1$. Figure 3.1 shows the generated data in black and the model estimates of the posterior in red. The posterior predictive distribution in the left panel shows that the MCMC sampling reasonably models the simulated data. We see in the right panel the the treatment effect is measured reasonably, but the estimates are noisy when considering that the effect is constant. Figure 3.2 shows where this noise comes from. Put simply, ϵ_i from Equation (3.6) is not a model input. The only inputs are the (x_i, y_i, z_i) triplets, so $\tau(x_i)$ is estimated as $\dfrac{y_i}{z_i}$ in practice.

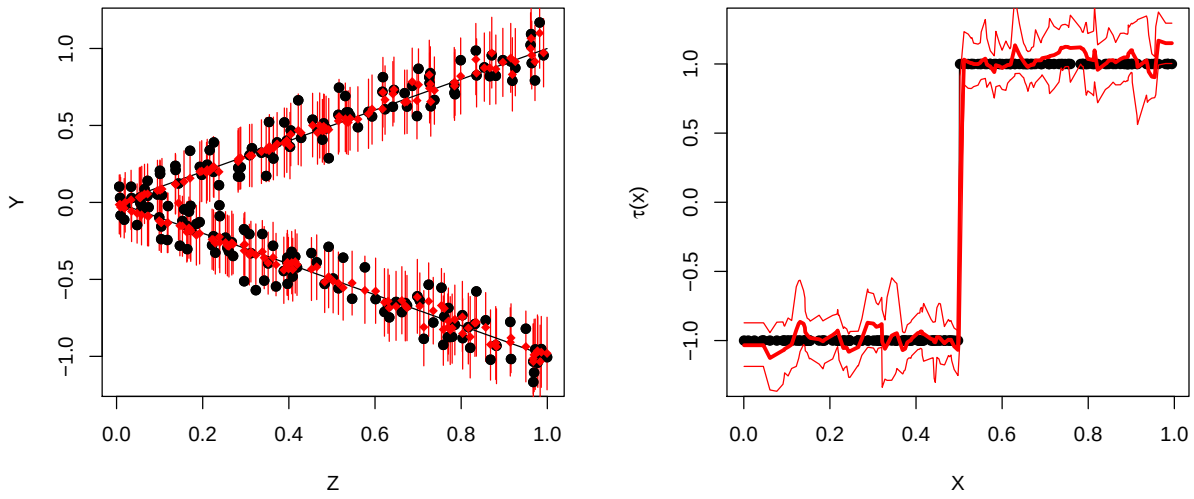
3.4.2 Generative Model with Medium Complication

We now add $\mu(x)$ back into the model, such that Equations (3.3) and (3.7) apply.

$$\mu(x) = \begin{cases} -1 & x \leq 0.75 \\ 1 & x > 0.75 \end{cases} \tag{3.8}$$

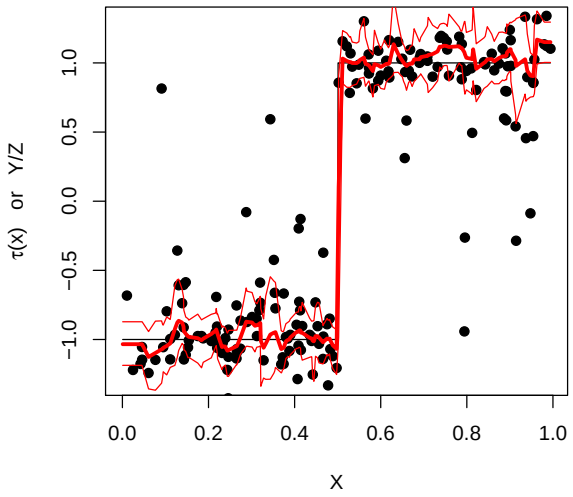
Figure 3.3 again shows the generated data in black and the model estimates of the posterior in red, and shows that the model fits step functions of the slope τ and intercept μ well.

Figure 3.1: Estimating a simple generative model



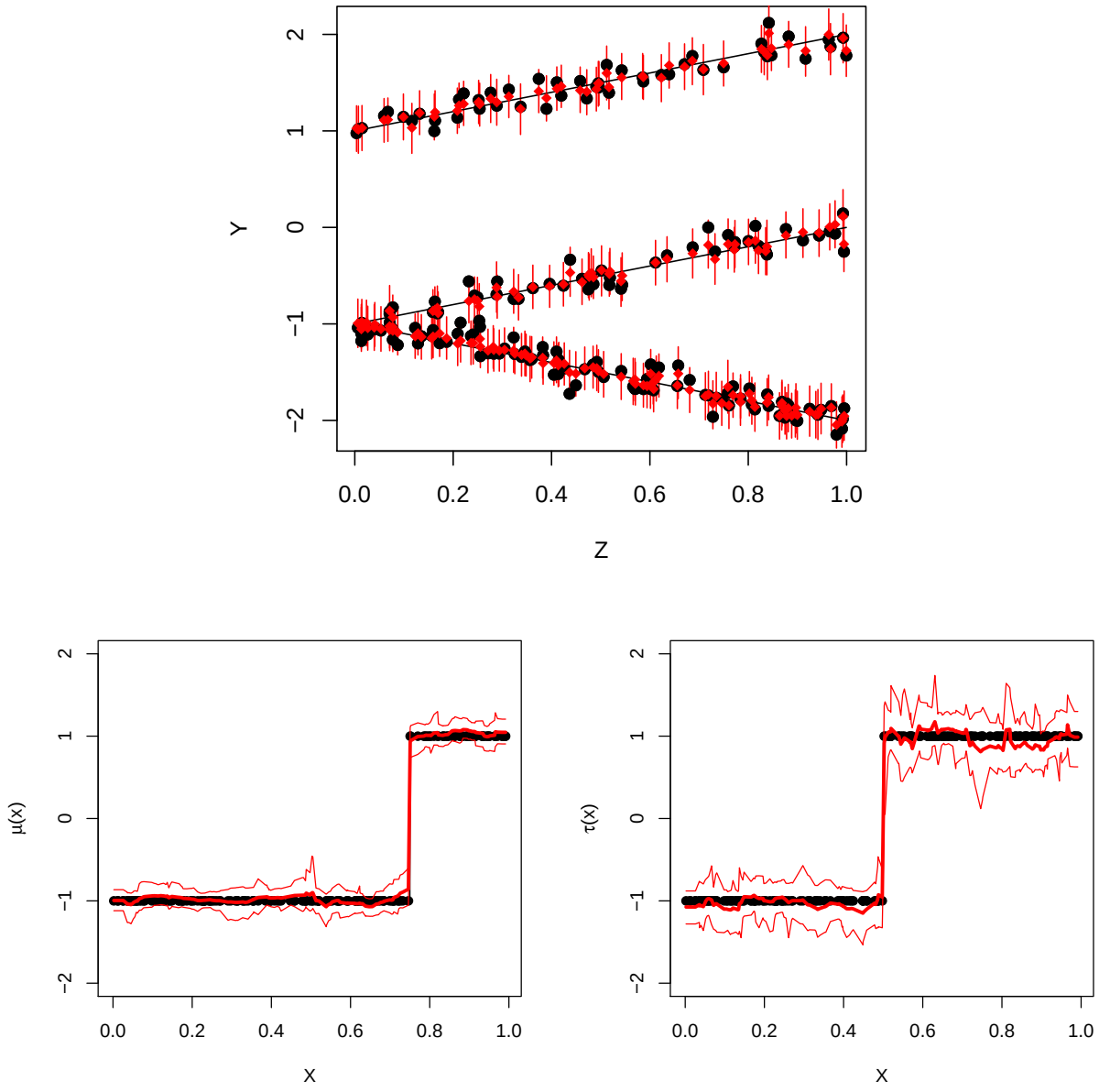
This figure shows the critical relationships of the simple model in Equations 3.6 and 3.7. The black components of the figure come from the generative model, while the red components come from the estimated posterior distribution. The black lines show the true relationship, and the black points are the generated data points. In the left panel, the red points and bars are, respectively, the posterior mean and 95% credible interval of the posterior predictive distribution for each y_i. In the right panel, the thick and thin red curves are, respectively, the posterior mean and 95% credible bands for the treatment effect slope τ.

Figure 3.2: Source of estimation noise



This figure shows the source of noise in the estimates of τ seen in the right panel of Figure 3.1. The black components of the figure come from the generative model, while the red components come from the estimated posterior distribution. The black lines show the true relationship, and the black points are the generated values of Y/Z. The thick and thin red curves are, respectively, the posterior mean and 95% credible bands for the treatment effect slope τ.

Figure 3.3: Estimating a more-complicated generative model



This figure shows the key relationships of the model in Equations (3.3), (3.7), and (3.8). The black components of the figure come from the generative model, while the red components come from the estimated posterior distribution. The black lines show the true relationship, and the black points are the generated data points. In the top panel, the red points and bars are, respectively, the posterior mean and 95% credible interval of the posterior predictive distribution for each y_i. In the bottom panels, the thick and thin red curves are, respectively, the posterior mean and 95% credible bands.

3.4.3 Generative Model with Nonlinear Treatment Effect

Not all phenomena will follow the same data generating process of our model. To experiment directly with this, we now generate from a logarithmic relationship:

$$y_i = \mu(x_i) + \tau(x_i)log(z_i) + \epsilon_i \tag{3.9}$$

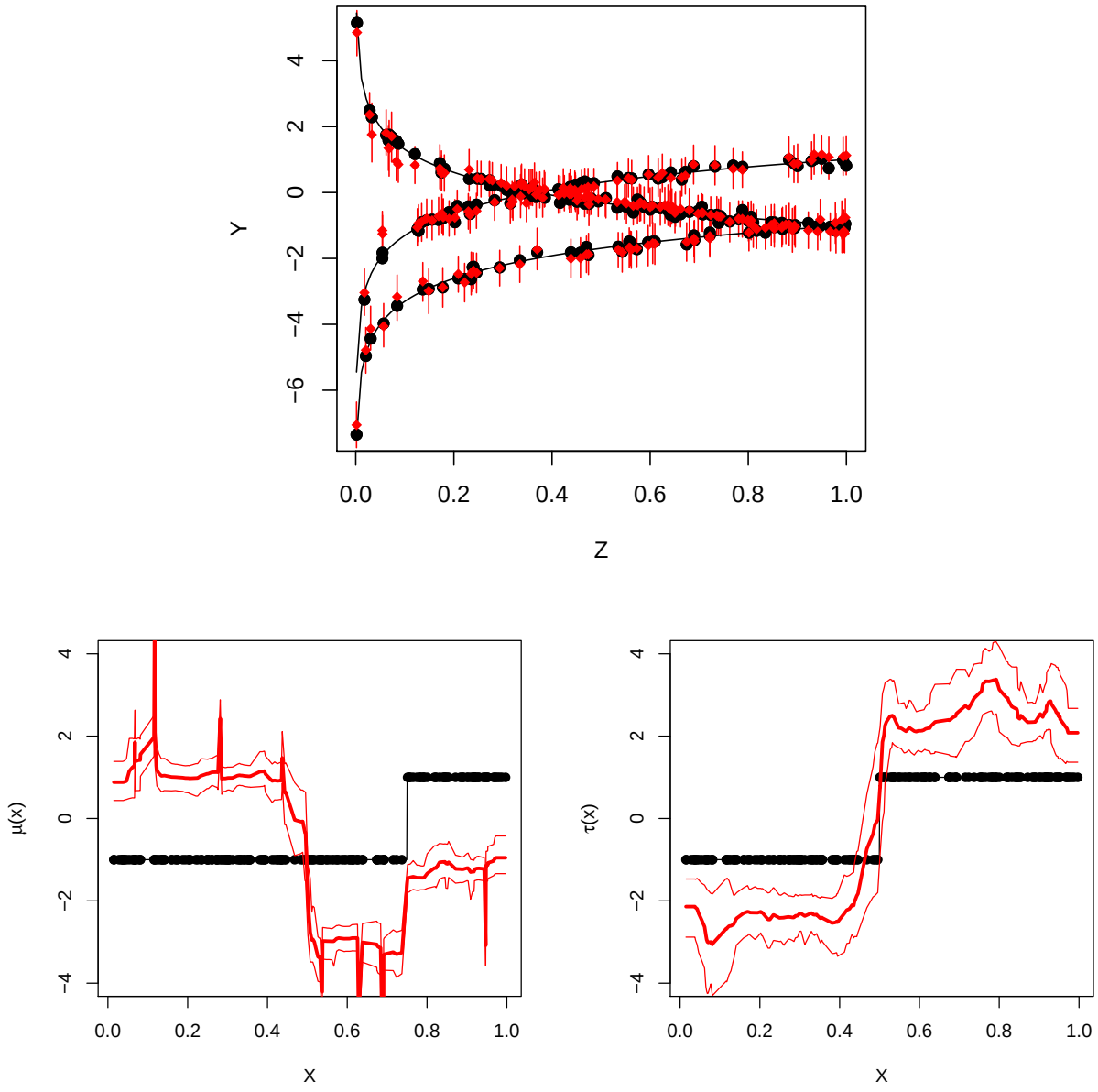
where μ is still defined by Equation (3.8) and τ by Equation (3.7). $n = 200$ and $\sigma = 0.1$ are also still used. The top panel of Figure 3.4 shows that, despite the data generating process being different from our model, the posterior distribution fits the observations well. However, the bottom panels illustrate that the lack of a linear relationship between y_i and z_i means we should no longer plan on recovering the original μ and τ.

Fortunately, the intent of our model is not recovering the underlying data generating process but to approximate it linearly. Figure 3.5 shows these approximations. Each color shows that the corresponding data point came from a different generative curve, in black. As per Equations (3.7) and (3.8), there are three defined regions in the support of x_i, which can also be seen in the posterior distribution of $\mu(x)$ in the bottom left panel of Figure 3.4. Figure 3.5 clearly shows shrinkage toward three dominant linear fits, with some lines straying from the dominant three. These strays reflect the spikes seen in the posterior distributions in the bottom panels of Figure 3.4. As we would expect, the linear fits are reasonable approximations to the true function in areas where the function is approximately linear (i.e. where $z_i > 0.2$). Otherwise, we recover the direction but not the magnitude of changes in the true function, due to the monotonicity of the true function within the three aforementioned regions.

3.5 Empirical Case Study: Financial Factors

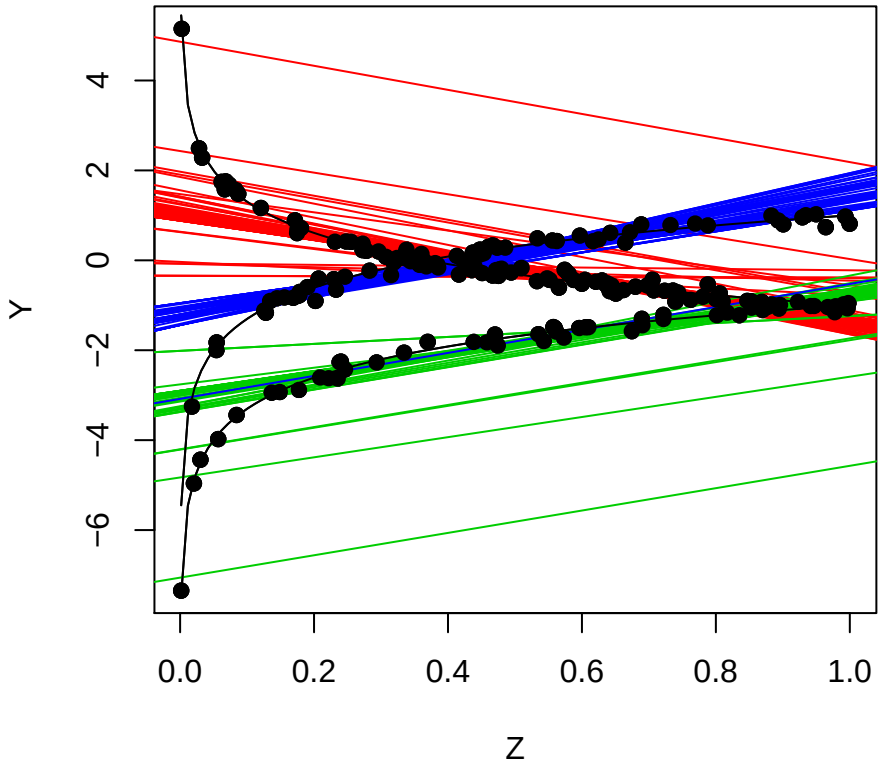
As detailed in Chapters 1 and 2, many variables in the finance literature are assumed to have reasonably-linear relationships with returns. Here, we return to the dataset used in

Figure 3.4: Estimating heterogenous, nonlinear relationships with linear models



This figure shows the key relationships of the model in Equation (3.9) . The black components of the figure come from the generative model, while the red components come from the estimated posterior distribution. The black lines shows the true relationship, and the black points are the generated data points. In the top panel, the red points and bars are, respectively, the posterior mean and 95% credible interval of the posterior predictive distribution for each y_i. In the bottom panels, the thick and thin red curves are, respectively, the posterior mean and 95% credible bands.

Figure 3.5: Estimated linear fits



This figure shows the linear fits resulting from the posterior mean intercept and slope for each point for the model in Equation (3.3). These linear fits are the colored lines, where each color represents a different section of X space as per Equations (3.7) and (3.8): red for $x < .5$, green for $x \in [.5, .75]$, blue for $x > .75$. The black lines show the true relationship from Equation (3.9), and the black points are the generated data points.

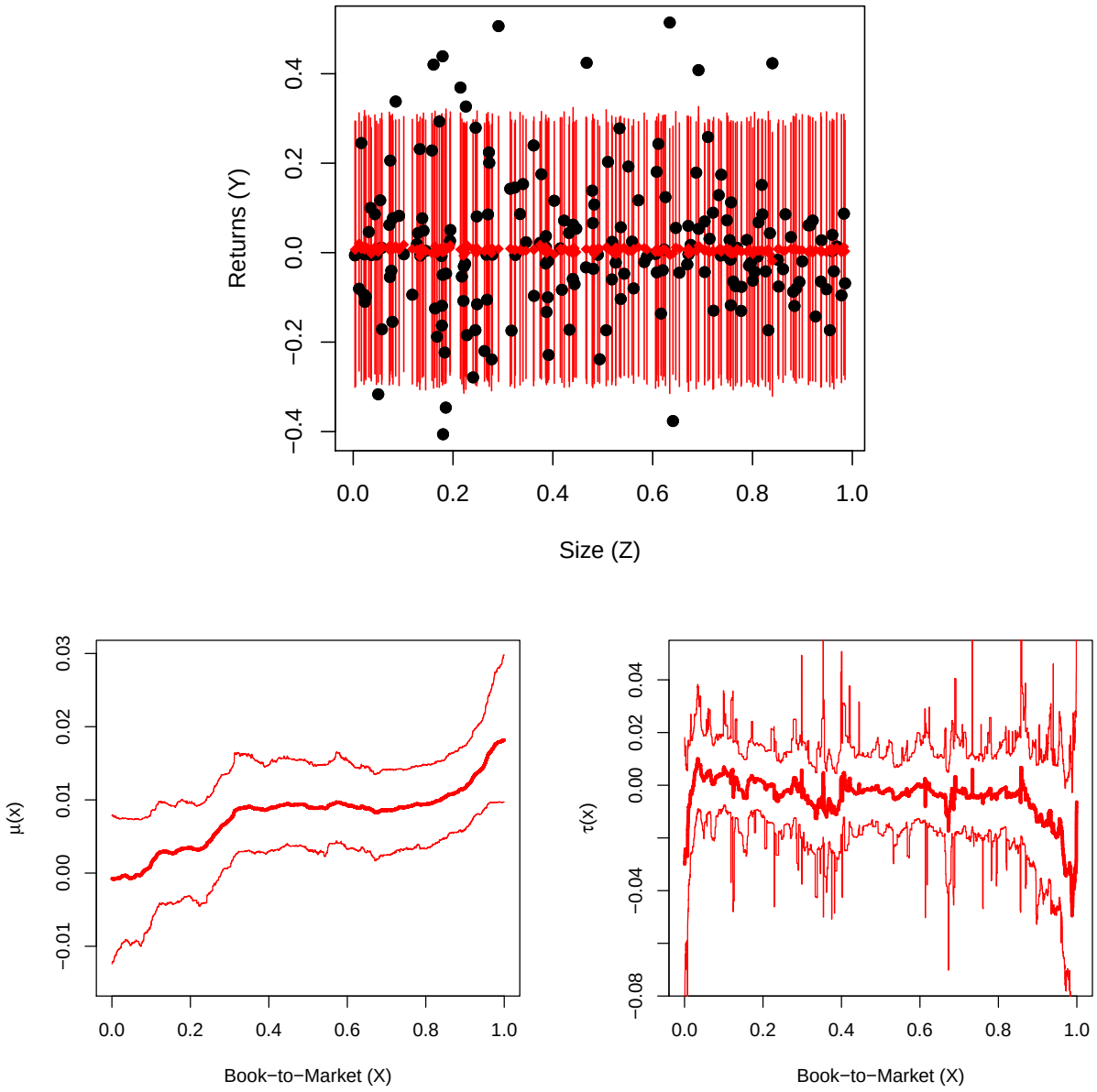
Chapter 2, from Freyberger et al. (2019). In particular, we look at size and book-to-market a la Fama and French (1993b). We subset the data to the 1963-1991 window to mimic Fama and French (1993b)'s dataset, though we randomly select 10% of firms each month to increase computation speed.

The top panel of Figure 3.6 shows what happens when, one, there's much more data, and two, the signal to noise ratio is much smaller. The posterior predictive fits appear to say there is no relationship, that a mean and variance of returns (y) is a sufficient description. But we know there is not much signal in the noise, and this is verified in the lower panels. The "intercept" $\mu(x)$ shows that for firms of the smallest size ($Z = 0$), returns increase with book-to-market, from 0% to 2% monthly, on average. This is not dissimilar to Fama and French (1993b), who show that from 1963-1991, the smallest quintile of firms' monthly return move from about 0.4% to about 1.0%. The story of size says that small firms have higher returns than large firms on average, thus the bottom right panel should show negative values. Indeed, we see that the posterior mean of $\tau(x)$ is usually negative, and its whole distribution is more negative for large values of book-to-market. This is also supported by Fama and French (1993b), who show that returns for firms in the lowest quintile of book-to-market have little change across sizes compared to the upper quintiles.

3.6 Conclusion and Future Work

Here, we have laid the groundwork for a powerful tool that detects heterogeneous linear relationships. Naturally, this could be simply expanded to include a variety of link functions, allowing for log-linear and logistic models. For more flexible function estimation, we can look to other BART adaptations, such as Starling et al. (2019), that use functions in the leaves of the regression trees instead of scalar values as we are using.

Figure 3.6: Estimating a returns as a linear function of size



This figure shows the key relationships of the model in Equation (3.3) applied to monthly stock excess returns. Only 200 of the data points are shown here as the black points, randomly chosen from the 125,239 total firm months, in order to clearly see the relationships. The red components come from the estimated posterior distribution. The top panel shows the 200 randomly chosen points, and the red points and bars are, respectively, the posterior mean and 95% credible interval of the posterior predictive distribution for each y_i. In the bottom panels, the thick and thin red curves are, respectively, the posterior mean and 95% credible bands.

Appendices

Appendix A

The Wishart DLM

This appendix comes from Fisher et al. (2019a) and describes our implementation of the W-DLM model of (West and Harrison, 1997, Section 16.4).

A.1 Basic Equations

For convenience, we reproduce here the key equations of the model. The W-DLM can be written as:

$$\boldsymbol{r}_t = \boldsymbol{B}_t' \boldsymbol{x}_{t-1} + \boldsymbol{v}_t \qquad\qquad \boldsymbol{v}_t | \boldsymbol{\Sigma}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_t) \tag{A.1}$$

where $\boldsymbol{B}_t$ is the $p \times q$ matrix of time-varying regression coefficients and $\boldsymbol{\nu_t}$ is a $q \times 1$ error vector, independent over time. The regression coefficients $\boldsymbol{B}_t$ vary over time according to pq random walk processes,

$$vec(\boldsymbol{B}_t) = vec(\boldsymbol{B}_{t-1}) + \boldsymbol{\omega}_t \qquad\qquad \boldsymbol{\omega}_t | \boldsymbol{\Sigma}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_t \otimes \boldsymbol{W}_t) \tag{A.2}$$

where $\boldsymbol{\omega}_t$ is a $pq \times 1$ vector of zero-mean normally distributed error terms. The initial conditions are given by

$$\begin{aligned} vec(\boldsymbol{B}_0) | \boldsymbol{\Sigma}_0, \mathcal{D}_0 &\sim \mathcal{N}(vec(\boldsymbol{M}_0), \boldsymbol{\Sigma}_0 \otimes \boldsymbol{C}_0) \\ \boldsymbol{\Sigma}_0 | \, \mathcal{D}_0 &\sim \mathcal{IW}(n_0, \boldsymbol{S}_0) \end{aligned} \tag{A.3}$$

A.2 Evolution

To begin the W-DLM forward filter, start at $t = 1$ such that the posterior distribution in step 1 below is the initial state of the filter based on $\mathcal{D}_{t-1} = \mathcal{D}_0$, the training dataset.

After the three steps of the filter below are fulfilled for time $t = 1$, repeat the steps for $t = 2, ..., T$, where T is the last time period in the data.

1. **Evolve Posterior of time $t-1$ to Prior of time t**

   Given the posterior distribution of the parameters at time $t-1$

$$\begin{aligned} vec(\boldsymbol{B}_{t-1})|\boldsymbol{\Sigma}_{t-1}, \mathcal{D}_{t-1} &\sim \mathcal{N}\left(vec(\boldsymbol{M}_{t-1}), \boldsymbol{\Sigma}_{t-1} \otimes \boldsymbol{C}_{t-1}\right) \\ \boldsymbol{\Sigma}_{t-1}|\mathcal{D}_{t-1} &\sim \mathcal{IW}(n_{t-1}, \boldsymbol{S}_{t-1}) \end{aligned} \tag{A.4}$$

   which we abbreviate as

$$\boldsymbol{B}_{t-1}, \boldsymbol{\Sigma}_{t-1}|\mathcal{D}_{t-1} \sim \mathcal{NIW}(\boldsymbol{M}_{t-1}, \boldsymbol{C}_{t-1}, n_{t-1}, \boldsymbol{S}_{t-1}). \tag{A.5}$$

   we evolve forward to create a prior for time t

$$\boldsymbol{B}_t, \boldsymbol{\Sigma}_t|\mathcal{D}_{t-1} \sim \mathcal{NIW}(\boldsymbol{M}_{t-1}, \hat{\boldsymbol{C}}_t, \hat{n}_t, \boldsymbol{S}_{t-1}). \tag{A.6}$$

   where, due to our choice of $\boldsymbol{W}_t$ in (1.8), $\hat{\boldsymbol{C}}_t = \frac{1}{\delta_\beta}\boldsymbol{C}_{t-1}$ and $\hat{n}_t = \delta_v n_{t-1}$, for chosen values of $\delta_\beta, \delta_v \in (0, 1]$.

2. **Forecast response variable at time t**

   As shown in equations (1.10)–(1.12), the predictive distribution of $\boldsymbol{r}_t$, based on time $t-1$ data, is given by

$$\boldsymbol{r}_t|\delta_\beta, \delta_v, \mathcal{D}_{t-1} \sim \mathcal{T}_{\hat{n}_t}\left(\boldsymbol{M}'_{t-1}\boldsymbol{x}_{t-1}, \quad \boldsymbol{S}_{t-1}(1 + \boldsymbol{x}'_{t-1}\hat{\boldsymbol{C}}_t\boldsymbol{x}_{t-1})\right). \tag{A.7}$$

   with mean and covariance matrix given by

$$\mathbb{E}[\boldsymbol{r}_t|\delta_\beta, \delta_v, \mathcal{D}_{t-1}] = \boldsymbol{M}'_{t-1}\boldsymbol{x}_{t-1} \tag{A.8}$$

$$cov[\boldsymbol{r}_t|\delta_\beta, \delta_v, \mathcal{D}_{t-1}] = \frac{\hat{n}_t}{\hat{n}_t - 2}\boldsymbol{S}_{t-1}(1 + \boldsymbol{x}'_{t-1}\hat{\boldsymbol{C}}_t\boldsymbol{x}_{t-1}). \tag{A.9}$$

3. **Update prior for time t into posterior for time t based on forecast error**

After observing $\boldsymbol{r}_t$ compute time t posterior distribution for $\boldsymbol{B}_t$ and $\boldsymbol{\Sigma}_t$:

$$\boldsymbol{B}_t, \boldsymbol{\Sigma}_t | D_t \sim \mathcal{NIW}(\boldsymbol{M}_t, \boldsymbol{C}_t, n_t, \boldsymbol{S}_t) \tag{A.10}$$

In particular, we have that

Posterior mean matrix	$\boldsymbol{M}_t = \boldsymbol{M}_{t-1} + \boldsymbol{a}_t \boldsymbol{e}_t'$	(A.11)
Posterior covariance matrix factor	$\boldsymbol{C}_t = \hat{\boldsymbol{C}}_t - q_t \boldsymbol{a}_t \boldsymbol{a}_t'$	(A.12)
Posterior degrees of freedom	$n_t = \hat{n}_t + 1$	(A.13)
Posterior residual covariance estimate	$\boldsymbol{S}_t = n_t^{-1}(\hat{n}_t \boldsymbol{S}_{t-1} + q_t^{-1} \boldsymbol{e}_t \boldsymbol{e}_t')$.	(A.14)

where

1-step ahead forecast error	$\boldsymbol{e}_t = \boldsymbol{r}_t - \boldsymbol{M}_{t-1}' \boldsymbol{x}_{t-1}$	(A.15)
1-step ahead coefficient variance factor	$q_t = 1 + \boldsymbol{x}_{t-1}' \hat{\boldsymbol{C}}_t \boldsymbol{x}_{t-1}$	(A.16)
Adaptive coefficient vector	$\boldsymbol{a}_t = q_t^{-1} \hat{\boldsymbol{C}}_t \boldsymbol{x}_{t-1}$	(A.17)

Appendix B

The Simultaneous Graphical DLM

This appendix comes from Fisher et al. (2019a) and describes our implementation of the SG-DLM of Gruber and West (2016).

B.1 Basic Equations

For convenience, we reproduce here the key equations of the SG-DLM. For $j = 1, ..., q$, we write

$$r_{jt} = \boldsymbol{x}'_{j,t-1}\boldsymbol{\beta}_{jt} + \boldsymbol{r}'_{<j,t}\boldsymbol{\gamma}_{<j,t} + \nu_{jt} \qquad \nu_{jt} \sim N(0, \sigma^2_{jt}) \tag{B.1}$$

where

$$\begin{pmatrix} \boldsymbol{\beta}_{jt} \\ \boldsymbol{\gamma}_{<j,t} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_{j,t-1} \\ \boldsymbol{\gamma}_{<j,t-1} \end{pmatrix} + \boldsymbol{\omega}_{jt} \qquad \boldsymbol{\omega}_{jt} \sim N(\mathbf{0}, \boldsymbol{W}_{jt}) \tag{B.2}$$

and the initial conditions are given by:

$$\begin{gathered} \begin{pmatrix} \boldsymbol{\beta}_{j0} \\ \boldsymbol{\gamma}_{<j,0} \end{pmatrix} \Bigg| \, \sigma^2_{j0}, \mathcal{D}_0 \sim \mathcal{N}\left(\boldsymbol{m}_{j0}, \frac{\sigma^2_{j0}}{s_{j0}}\boldsymbol{C}_{j0}\right) \\ \sigma^{-2}_{j0} | \mathcal{D}_0 \sim \mathcal{G}\left(\frac{n_{j0}}{2}, \frac{n_{j0}s_{j0}}{2}\right) \end{gathered} \tag{B.3}$$

B.2 Evolution

To begin the SG-DLM forward filter, begin with $t = 1$ such that the posterior distribution in step 1 below is the initial state of the filter based on $\mathcal{D}_{t-1} = \mathcal{D}_0$, the training dataset. After the three steps of the filter are fulfilled for time $t = 1$, repeat the steps for $t = 2, ..., T$, where T is the last time period in the data.

1. **Evolve Posterior of time $t-1$ to Prior of time t**

   Given the Posterior at time $t-1$

$$\begin{pmatrix} \boldsymbol{\beta}_{j,t-1} \\ \boldsymbol{\gamma}_{<j,t-1} \end{pmatrix} \Bigg| \sigma^2_{j,t-1}, \mathcal{D}_{t-1} \sim \mathcal{N}\left(\boldsymbol{m}_{j,t-1}, \quad \frac{\sigma^2_{j,t-1}}{s_{j,t-1}} \boldsymbol{C}_{j,t-1}\right) \tag{B.4}$$

$$\sigma^{-2}_{j,t-1} | \mathcal{D}_{t-1} \sim \mathcal{G}\left(\frac{n_{j,t-1}}{2}, \quad \frac{n_{j,t-1} s_{j,t-1}}{2}\right), \tag{B.5}$$

   which we abbreviate as

$$\begin{pmatrix} \boldsymbol{\beta}_{j,t-1} \\ \boldsymbol{\gamma}_{<j,t-1} \end{pmatrix}, \sigma^2_{j,t-1} \Bigg| \mathcal{D}_{t-1} \sim \mathcal{NG}(\boldsymbol{m}_{j,t-1}, \boldsymbol{C}_{j,t-1}, n_{j,t-1}, s_{j,t-1}), \tag{B.6}$$

   we evolve it to create a prior for time t

$$\begin{pmatrix} \boldsymbol{\beta}_{j,t} \\ \boldsymbol{\gamma}_{<j,t} \end{pmatrix}, \sigma^2_{j,t} \Bigg| \mathcal{D}_{t-1} \sim \mathcal{NG}(\boldsymbol{m}_{j,t-1}, \hat{\boldsymbol{C}}_{jt}, \hat{n}_{jt}, s_{j,t-1}) \tag{B.7}$$

   where, due to our choice of $\boldsymbol{W}_t$ in (1.25), $\hat{\boldsymbol{C}}_{jt} = \begin{bmatrix} \boldsymbol{C}_{\beta\beta_{j,t-1}}/\delta_{\beta j} & \boldsymbol{C}_{\beta\gamma_{j,t-1}} \\ \boldsymbol{C}_{\gamma\beta_{j,t-1}} & \boldsymbol{C}_{\gamma\gamma_{j,t-1}}/\delta_{\gamma j} \end{bmatrix}$ and $\hat{n}_{jt} = \delta_{vj} n_{j,t-1}$, where $\boldsymbol{C}_{\beta\beta_{j,t-1}}$ and $\boldsymbol{C}_{\gamma\gamma_{j,t-1}}$ are, respectively, the covariance matrix factors for $\boldsymbol{\beta}_{j,t-1}$ and $\boldsymbol{\gamma}_{j,t-1}$.

2. **Forecast response variable at time t**

   We calculate the moments of forecast returns one asset at a time, according to the order of dependence. Similarly-derived forecasting steps for this type of model can be found Zhao et al. (2016), Appendix B. As it does not depend on other assets' returns, the forecast for the first asset is given by

$$r_{1t} | \boldsymbol{\delta}_j, \mathcal{D}_{t-1} \sim \mathcal{T}_{\hat{n}_{1t}}\left(\boldsymbol{x}'_{1,t-1}\boldsymbol{m}_{1,t-1}, \boldsymbol{x}'_{1,t-1}\hat{\boldsymbol{C}}_{1t}\boldsymbol{x}_{1,t-1} + s_{1,t-1}\right) \tag{B.8}$$

   with mean and variance that are equal to:

$$\mathbb{E}[r_{1t} | \boldsymbol{\delta}_j, \mathcal{D}_{t-1}] = \boldsymbol{x}'_{1,t-1}\boldsymbol{m}_{1,t-1} \tag{B.9}$$

$$Var[r_{1t} | \boldsymbol{\delta}_j, \mathcal{D}_{t-1}] = \frac{\hat{n}_{1t}}{\hat{n}_{1t}-2}(\boldsymbol{x}'_{1,t-1}\hat{\boldsymbol{C}}_{1t}\boldsymbol{x}_{1,t-1} + s_{1,t-1}) \tag{B.10}$$

where $\boldsymbol{\delta}_j = (\delta_{\beta j}, \delta_{\gamma j}, \delta_{vj})$. Now, all other assets' forecast moments can be found sequentially ($j = 2, ..., q$). Similarly, their conditional distributions follow Student's t-distribution, with predictive moments given by

$$\mathbb{E}[r_{jt}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}] = \boldsymbol{x}'_{j,t-1}\boldsymbol{m}_{\beta_{j,t-1}} + \mathbb{E}[\boldsymbol{r}_{<j,t}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}]'\boldsymbol{m}_{\gamma_{<j,t-1}}, \tag{B.11}$$

$$\begin{aligned} Var[r_{jt}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}] = \frac{\hat{n}_{jt}}{\hat{n}_{jt}-2}\left\{tr\left(\hat{\boldsymbol{C}}_{\gamma_{<j,t}} Cov[\boldsymbol{r}_{<j,t}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}]\right) + c_{jt} + s_{j,t-1}\right\} \\ + \boldsymbol{m}'_{\gamma_{<j,t-1}} Cov[\boldsymbol{r}_{<j,t}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}]\boldsymbol{m}_{\gamma_{<j,t-1}} \end{aligned} \tag{B.12}$$

and

$$Cov[r_{jt}, \boldsymbol{r}_{<j,t}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}] = \boldsymbol{m}'_{\boldsymbol{\gamma}_{<j,t-1}} Cov[\boldsymbol{r}_{<j,t}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}] \tag{B.13}$$

where $\mathbb{E}[r_{<jt}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}]$ and $Cov[\boldsymbol{r}_{<j,t}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}]$ are known, $tr()$ stands for the trace of a matrix, and

$$c_{jt} = \begin{pmatrix} \boldsymbol{x}_{j,t-1} \\ \mathbb{E}[\boldsymbol{r}_{<j,t}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}] \end{pmatrix}' \hat{\boldsymbol{C}}_{jt} \begin{pmatrix} \boldsymbol{x}_{j,t-1} \\ \mathbb{E}[\boldsymbol{r}_{<j,t}|\boldsymbol{\delta}_j, \mathcal{D}_{t-1}] \end{pmatrix}. \tag{B.14}$$

3. **Update prior for time t into posterior for time t based on forecast error**

After observing $\boldsymbol{r}_t$, time t posterior distribution for $\boldsymbol{\beta}_{j,t}$, $\boldsymbol{\gamma}_{<j,t}$, and $\sigma^2_{j,t}$ ($j = 1, ..., q$) are given by

$$\begin{pmatrix} \boldsymbol{\beta}_{j,t} \\ \boldsymbol{\gamma}_{<j,t} \end{pmatrix}, \sigma^2_{j,t} \Big| \mathcal{D}_t \sim \mathcal{NG}(\boldsymbol{m}_{j,t}, \boldsymbol{C}_{j,t}, n_{j,t}, s_{j,t}). \tag{B.15}$$

In particular, we have that

Posterior mean vector	$\boldsymbol{m}_{jt} = \boldsymbol{m}_{j,t-1} + \boldsymbol{a}_{jt}e_{jt}$	(B.16)
Posterior covariance matrix factor	$\boldsymbol{C}_{jt} = (\hat{\boldsymbol{C}}_{jt} - \boldsymbol{a}_{jt}\boldsymbol{a}'_{jt}q_{jt})z_{jt}$	(B.17)
Posterior degrees of freedom	$n_{jt} = \hat{n}_{jt} + 1$	(B.18)
Posterior residual variance estimate	$s_{jt} = z_{jt}s_{j,t-1}.$	(B.19)

where

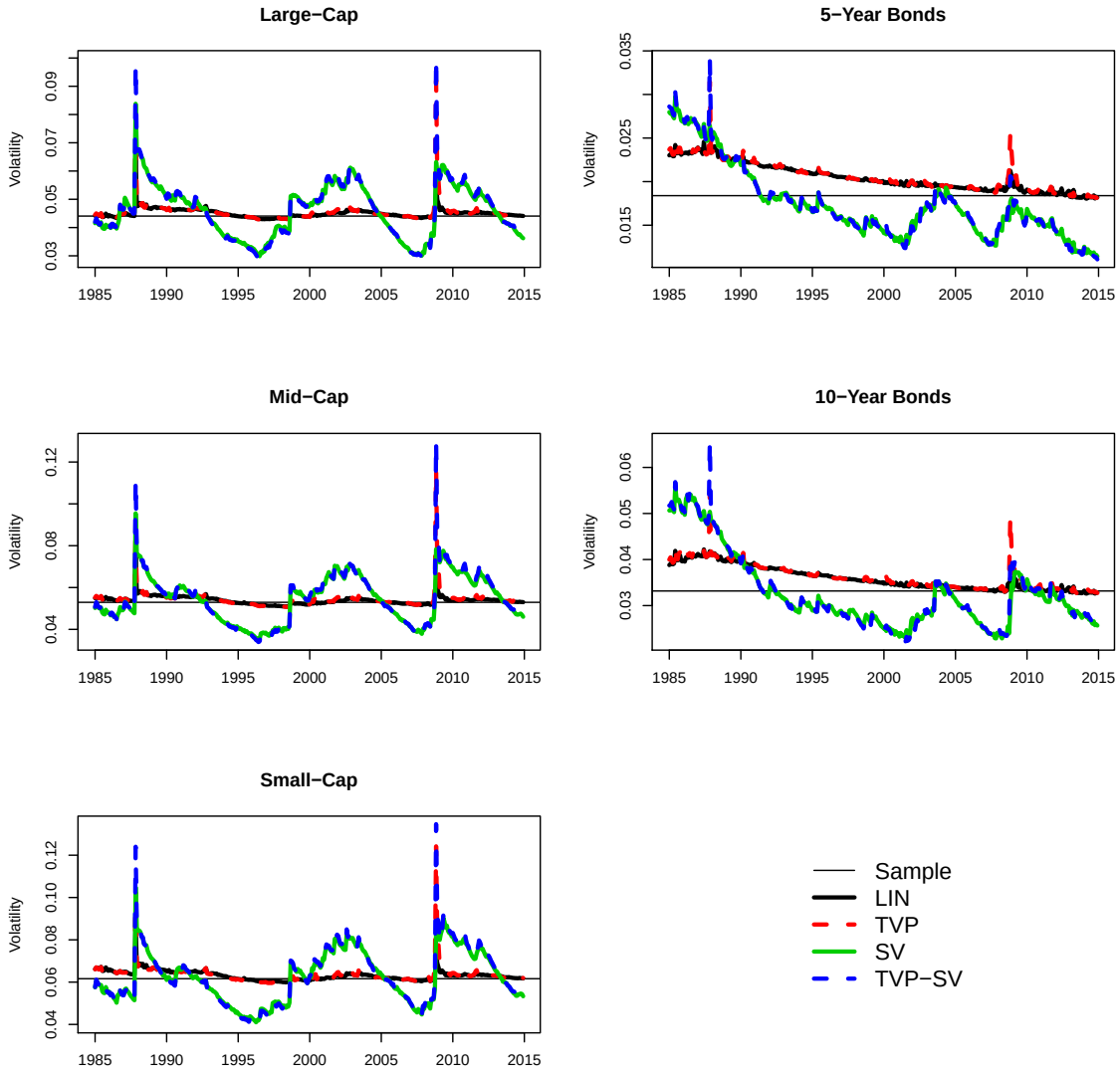
1-step ahead forecast error	$e_{jt} = r_{jt} - \begin{pmatrix} \boldsymbol{x}_{j,t-1} \\ \boldsymbol{r}_{<j,t} \end{pmatrix}' \boldsymbol{m}_{j,t-1}$	(B.20)
1-step ahead forecast variance factor	$q_{jt} = s_{j,t-1} + \begin{pmatrix} \boldsymbol{x}_{j,t-1} \\ \boldsymbol{r}_{<j,t} \end{pmatrix}' \hat{\boldsymbol{C}}_{jt} \begin{pmatrix} \boldsymbol{x}_{j,t-1} \\ \boldsymbol{r}_{<j,t} \end{pmatrix}$	(B.21)
Adaptive coefficient vector	$\boldsymbol{a}_{jt} = \hat{\boldsymbol{C}}_{jt} \begin{pmatrix} \boldsymbol{x}_{j,t-1} \\ \boldsymbol{r}_{<j,t} \end{pmatrix} / q_{jt}$	(B.22)
Volatility update factor	$z_{jt} = (\hat{n}_{jt} + e_{jt}^2/q_{jt})/(\hat{n}_{jt} + 1)$	(B.23)

Appendix C

Additional Results on the Wishart DLM

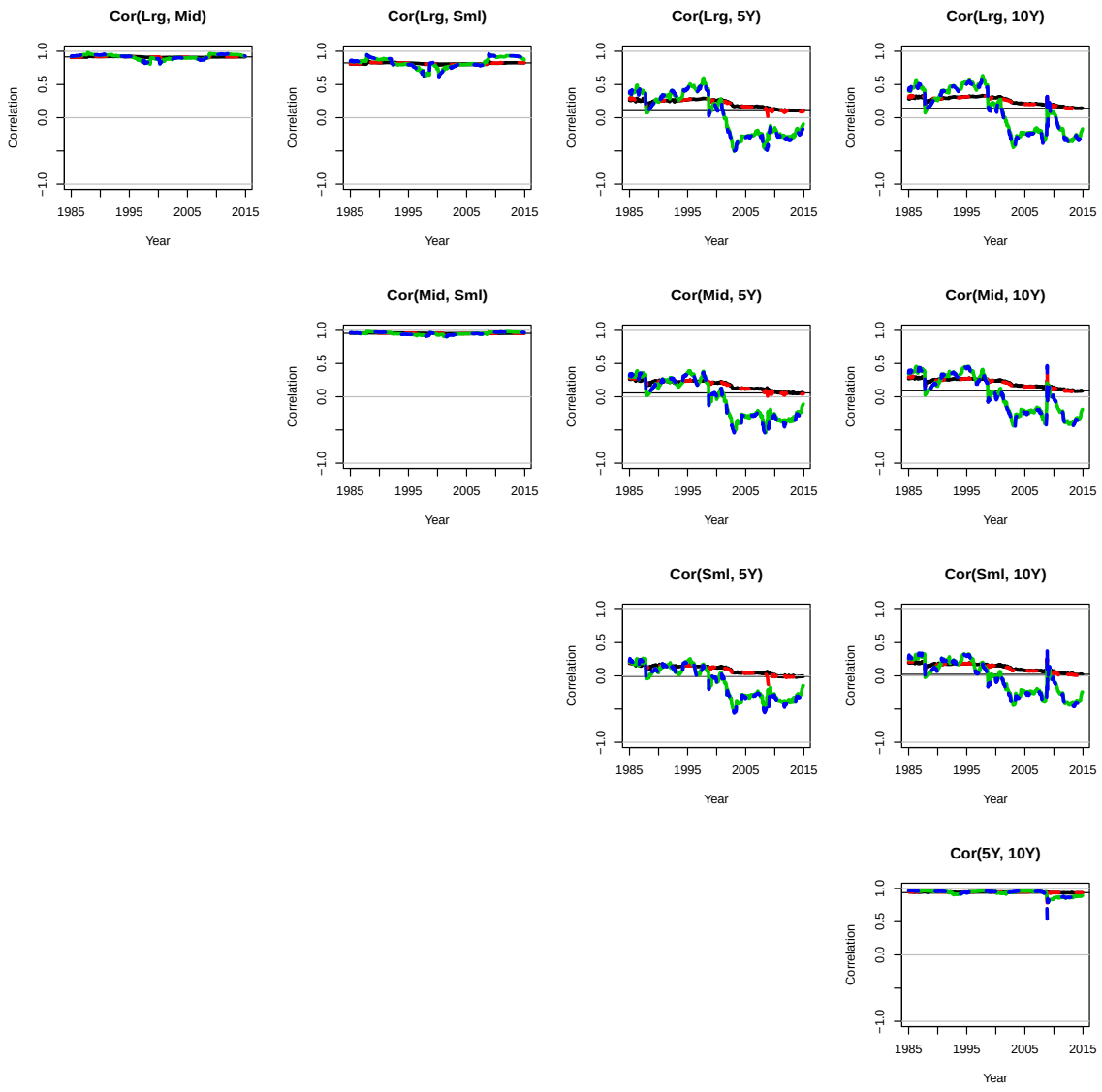
This appendix is from Fisher et al. (2019a) and presents additional figures on the performance of the Wishart DLM (W-DLM).

Figure C.1: Time series of predicted volatilities for W-DLM models



The figure shows the time-series of predicted volatilities of expected excess returns for the four variants of the W-DLM score-based model combinations, namely LIN, TVP, SV, and TVP-SV. Each panel represents a different asset, as labeled. Note that the scales of the vertical axes are different for each asset in order to compare patterns of change over time, as opposed comparing the magnitude of volatilities across assets. The solid black line represents the LIN model; the dotted red line tracks the TVP model; the solid green line depicts the SV model, while the blue dotted line displays the TVP-SV model. In each panel we also display, as a reference, the level of the unconditional standard deviation of each asset, computed over the whole evaluation period, January 1985 – December 2014.

Figure C.2: Time series of predicted correlations for W-DLM models



The figure shows the time-series of predicted correlations of expected excess returns for the four variants of the W-DLM score-based model combinations, namely LIN, TVP, SV, and TVP-SV. Each panel represents a different pair of asset returns, as labeled. The solid black line represents the LIN model; the dotted red line tracks the TVP model; the solid green line depicts the SV model, while the blue dotted line displays the TVP-SV model. In each panel we also display, as a reference, the level of the unconditional correlation between each pair of asset returns, computed over the whole evaluation period, January 1985 – December 2014.

Appendix D

Description of Characteristics Data

This appendix is from Fisher et al. (2019b) and describes the data from Freyberger et al. (2019), along with the direction of monotonicity suggested by the literature. Papers in the literature were referenced for established directions of monotonic relationships. The relationship is classified as either positive (monotonic increasing), negative, or unclear. The unclear category contains non-monotonic variables, variables whose literature is undecided on the direction, as well as variables whose relationship with returns is unclear.

Table D.1: Firm Characteristics and references for direction of relationship with returns

Variable	Description	Papers	Monotonic Direction
a2me	Total assets to size	Bhandari (1988)	Unclear
at	Total assets	Gandhi and Lustig (2015)	Unclear
ato	Asset turnover: Sales to lagged net operating assets	Soliman(2008)	Positive
beme	Book to market ratio	Rosenberg, Reid, and Lanstein (1985); Davis, Fama and French (2000); Stattman (1980); Rosenberg et al. (1985); and Fama and French (1992)	Positive
beta	CAPM Beta	Frazzini and Pedersen (2014); Black et al. (1972); Fama and MacBeth (1973); Fama and French (1992); Fama and French (2006)	Negative
c	Cash to total assets	Palazzo (2012)	Positive

Continued on next page

Table D.1 – continued from previous page

Variable	Description	Papers	Monotonic Direction
cto	Sales to lagged total assets	Haugen and Baker (1996)	Unclear
d2a	Depreciation and amortization (DP) to total assets (AT)	Gorodnichenko and Weber (2016)	Unclear
dpi2a	Change in PP&E and inventory over lagged assets (AT)	Lyandres, Sun, and Zhang (2008)	Negative
e2p	Earnings to price	Basu (1983)	Positive
fc2y	Fixed costs to sales	D'Acunto, Liu, Pflueger, and Webern (2016)	Unclear
free_cf	Free cash flow to book equity	Hou et al. 2011	Positive
idio_vol	Idiosyncratic volatility from Fama-French 3 factor model	Ang, Hodrick, Xing, and Zhang (2006)	Negative
investment	Percent change in total assets	Cooper, Gulen, and Schill (2008)	Negative
lev	Leverage	Lewellen (2015); Bhandari (1988); Fama and French (1992)	Positive
lme	Size: market equity defined as stock price times shares outstanding	Fama and French (1992); Banz (1981); Lewellen (2015); Fama and French (2008)	Negative
lturnover	Volume to shares outstanding (turnover)	Datar, Naik, and Radcliffe (1998); Lee and Swaminathan (2000)	Negative
noa	Net-operating assets over lagged assets (AT)	Hirshleifer, Hou, Teoh, and Zhang (2004)	Negative
oa	Operating accruals	Sloan (1996)	Unclear
ol	Costs of goods sold + SG&A to total assets	Novy-Marx (2011)	Positive
pcm	Price-to-cost margin: Sales minus costs of goods sold to sales	Bustamante and Donangelo (2016); Gorodnichenko and Weber (2016); D'Acunto, Liu, Pfluueger, and Weber (2017)	Positive
pm	Profit margin: OI after depreciation over sales	Soliman (2008)	Positive

Continued on next page

Table D.1 – continued from previous page

Variable	Description	Papers	Monotonic Direction
prof	Profitability: Gross profitability over BE	Ball, Gerakos, Linnainmaa, and Nikolaev (2015); Lewellen (2015)	Positive
q	Tobin's Q		Unclear
r_{12-2}	Momentum	Fama and French (1996)	Positive
r_{12-7}	Intermediate momentum	Novy-Marx (2012)	Positive
r_{2-1}	Short-term reversal		Unclear
r_{36-13}	Long-term reversal	De Bondt and Thaler (1985)	Unclear
rel_to_high_price	Price to 52-week-high price	George and Hwang (2004)	Positive
rna	Return on net operating assets: OI after depreciation to lagged net operating assets	Soliman (2008)	Positive
roa	Return on assets: Income before extraordinary items to lagged AT	Balakrishnan, Bartov, and Faurel (2010)	Positive
roe	Return on equity: Income before extraordinary items to lagged BE	Haugen and Baker (1996)	Positive
s2p	Sales to price	Lewellen (2015); Fama and French (1992); Lakonishok et al. (1994)	Positive
sga2m	Expenses-to-sales: ratio of expenses (XSGA) to net sales (SALE)		Unclear
spread_mean	Average daily bid-ask spread	Chung and Zhang (2014)	Unclear
suv	Standard unexplained volume	Garfinkel (2009)	Unclear

Appendix E

Statistical Formulation and Computation of Additive Monotonic Splines Model

This appendix is similar to an appendix from Fisher et al. (2019b) and describes the formulation and computation of the additive monotonic splines model in Section 2.2. The text here is taken from and similar to the appendices in.

E.1 Model Summary

We model the vector of returns$\boldsymbol{r}_t$ of n_t firms using the vector of unknowns $\boldsymbol{\theta}_t$,

$$\begin{aligned}
\boldsymbol{r}_t|\boldsymbol{\theta}_t &\sim N\left(\alpha_t \mathbf{1}_{n_t} + \sum_{k=1}^{K} f_{kt}(\boldsymbol{x}_{k,t-1}),\ \sigma_t^2 I_n\right) \\
f_{kt}(\boldsymbol{x}_{k,t-1}) &= X_{k,t-1}\boldsymbol{\beta}_{kt} = X_{k,t-1}L^{-1}L\boldsymbol{\beta}_{kt} = W_{kt}L\boldsymbol{\beta}_{kt} = W_{kt}\boldsymbol{\gamma}_{kt} \\
\alpha_t &\sim N(0, 10^{-2}) \\
\sigma_t^2 &\sim U(0, 10^3) \\
(\gamma_{jkt}|I_{jkt} = 1, \cdot) &\sim N_+(0, c_k\sigma_t^2) \\
(\gamma_{jkt}|I_{jkt} = 0) &= 0 \\
I_{jkt} &\sim Bn(p_{jk} = 0.2)
\end{aligned}$$

where L describes the coefficients needed for monotonic constraints to hold. For a nonnegative spline, $0 \leq L\boldsymbol{\beta}_{kt}$ must hold. For nonpositive splines, the inequality flips. The coefficients of L are discussed in the next section.

E.2 Spline Conditions

Without loss of generality, we want $0 \leq f'(x)$ for all $x \in [-0.5, 0.5]$. With these splines, we get $\grave{m} + \acute{m} + 3$ constraints that bind:

$$0 \leq f'(-0.5) = \beta_1 + 2\beta_2(-0.5) + 2\beta_3(-0.5 - \grave{x}_1) + ... + 2\beta_{\grave{m}+2}(-0.5 - \grave{x}_{\grave{m}})$$

$$0 \leq f'(\grave{x}_{\grave{m}}) = \beta_1 + 2\beta_2(\grave{x}_{\grave{m}}) + 2\beta_3(\grave{x}_{\grave{m}} - \grave{x}_1) + ... + 2\beta_{\grave{m}+1}(\grave{x}_{\grave{m}} - \grave{x}_{\grave{m}-1})$$

$$\vdots$$

$$0 \leq f'(\grave{x}_2) = \beta_1 + 2\beta_2(\grave{x}_2) + 2\beta_3(\grave{x}_2 - \grave{x}_1)$$

$$0 \leq f'(\grave{x}_1) = \beta_1 + 2\beta_2(\grave{x}_1)$$

$$0 \leq f'(0) = \beta_1$$

$$0 \leq f'(\acute{x}_1) = \beta_1 + 2\beta_{\grave{m}+3}(\acute{x}_1)$$

$$0 \leq f'(\acute{x}_2) = \beta_1 + 2\beta_{\grave{m}+3}(\acute{x}_2) + 2\beta_{\grave{m}+4}(\acute{x}_2 - \acute{x}_1)$$

$$\vdots$$

$$0 \leq f'(\acute{x}_{\acute{m}}) = \beta_1 + 2\beta_{\grave{m}+3}(\acute{x}_{\acute{m}}) + 2\beta_{\grave{m}+4}(\acute{x}_{\acute{m}} - \acute{x}_1) + ... + 2\beta_{\grave{m}+\acute{m}+2}(\acute{x}_{\acute{m}} - \acute{x}_{\acute{m}-1})$$

$$0 \leq f'(0.5) = \beta_1 + 2\beta_{\grave{m}+3}(0.5) + 2\beta_{\grave{m}+4}(0.5 - \acute{x}_1) + ... + 2\beta_{\grave{m}+\acute{m}+3}(0.5 - \acute{x}_{\acute{m}})$$

which can be vectorized as a system of $\grave{m} + \acute{m} + 3$ linear inequalities, and these inequalities serve as our monotonicity conditions.

E.3 The MCMC Sampler

To sample all parameters at time $\tau \in \{1, ..., T\}$, iterate through the following, conditional upon the most recent draws of other parameters:

1. Draw $\alpha_\tau \sim N(m_\alpha, v_\alpha)$
   - $m_\alpha = \frac{v_\alpha}{\sigma^2} \sum_{t=1}^{\tau} \omega_t \mathbf{1}'_{n_t} \left(\boldsymbol{r}_t - \sum_{k=1}^{K} W_{kt} \boldsymbol{\gamma}_{k\tau} \right)$ and
   - $v_\alpha = \left(\frac{1}{\sigma^2} \sum_{t=1}^{\tau} \omega_t n_t + \frac{1}{10^{-2}} \right)^{-1}$.

2. Draw $\sigma_\tau^2 \sim IG(a_\sigma, b_\sigma)$, where

   - $a_\sigma = \frac{1}{2}\left(\sum_{t=1}^{\tau} n_t\omega_t + \sum_{j=1}^{m}\sum_{k=1}^{K} I_{jk\tau}\right) - 1$ and
   - $b_\sigma = \frac{1}{2}\left(\sum_{t=1}^{\tau} \omega_t \boldsymbol{e}_t'\boldsymbol{e}_t + \sum_{j=1}^{m}\sum_{k=1}^{K} \frac{\gamma_{jk\tau}^2}{c_k}\right)$ for the residual $\boldsymbol{e}_t = \boldsymbol{r}_t - \alpha_\tau \mathbf{1}_{n_t} - \sum_{k=1}^{K} W_{kt}\boldsymbol{\gamma}_{k\tau}$.

3. For coefficients $j = 1, ..., m+2$ and characteristics $k = 1, ..., K$:

   (a) Draw $I_{jk\tau} \sim Bernoulli(p_{jk\tau}^*)$ where

   - $p_{jk\tau}^* = \frac{\hat{p}_{jk\tau}}{\hat{p}_{jk\tau} + (1 - p_{jk})}$
   - $\hat{p}_{jk\tau} = 2p_{jk}c_k^{-\frac{1}{2}} v_{\gamma_{jk\tau}}^{\frac{1}{2}} exp\left\{\frac{1}{2\sigma^2 v_{\gamma_{jk\tau}}} m_{\gamma_{jk\tau}}^2\right\}\left[1 - \Phi\left(0 \,\middle|\, m_{\gamma_{jk\tau}},\ \sigma^2 v_{\gamma_{jk\tau}}\right)\right]$
   - $m_{\gamma_{jk\tau}} = v_{\gamma_{jk\tau}} \sum_{t=1}^{\tau} \omega_t \boldsymbol{e}_{(jk)t}' \boldsymbol{w}_{jkt}$
   - $v_{\gamma_{jk\tau}} = \left(\sum_{t=1}^{\tau} \omega_t \boldsymbol{w}_{jkt}'\boldsymbol{w}_{jkt} + \frac{1}{c_k}\right)^{-1}$
   - $\boldsymbol{e}_{(jk)t} = \boldsymbol{r}_t - \alpha_\tau \mathbf{1}_{n_t} - \sum_{\ell \neq k} W_{\ell t}\boldsymbol{\gamma}\boldsymbol{\tau}_\ell - \sum_{\ell \neq j} \boldsymbol{w}_{\ell kt}\gamma_{\ell k\tau}$, the residual assuming $\gamma_{jk\tau} = 0$

   (b) If $I_{jk\tau} = 1$ then draw $\gamma_{jk\tau} \sim N_+\left(m_{\gamma_{jk\tau}},\ \sigma^2 v_{\gamma_{jk\tau}}\right)$, else $\gamma_{jk\tau} = 0$.

Appendix F

Posterior calculations when using BART priors

In this Appendix, I describe the BART prior of Chipman et al. (2010) and some important posterior derivations of the model in Chapter 3.

F.1 BART Model

BART uses a sum of trees, namely

$$y_i = \sum_{j=1}^{m} g(\boldsymbol{x}_i; T_j, M_j) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2).$$

The full model is as follows, where θ_i is my personal addition for added clarity.

$$\begin{aligned} y_i|T, M, \sigma^2 &\sim N\left(\theta_i, \sigma^2\right) \\ \theta_i &= \sum_{j=1}^{m} g(\boldsymbol{x}_i; T_j, M_j) \\ \mu_{ij}|T_j &\sim N(\mu_\mu, \sigma_\mu^2) \\ T_j &\sim p(T_j), \text{ (Chipman et al., 2010)} \\ \sigma^2 &\sim \nu\lambda\chi_\nu^{-2} = IG(\nu/2, \nu\lambda/2) \end{aligned}$$

Instead of choosing μ_μ, Chipman et al. (2010) suggest transforming y such that $y_{max} = 0.5$, $y_{min} = -0.5$, and let $\mu_{ij} \sim N(0, \sigma_\mu^2)$ with $\sigma_\mu = \frac{1}{2k\sqrt{m}}$.

Hyperparameters ν, λ need to be tuned. Typically, $\nu \in [3, 10]$ is chosen, and then choose λ such that $P(\sigma < \hat{\sigma}) = q$, where $q \in (0, 1)$ is large e.g. 0.17, 0.9, 0.99. $\hat{\sigma}$ is a naive overestimate of σ, like the least squares standard error or the standard deviation of y.

According to Chipman et al. (2010), $p(T)$ is a "tree-generating stochastic process." Beginning with a single node tree, a terminal node η is split with probability $p_{SPLIT}(\eta, T) = \alpha(1 + d_\eta)^{-\beta}$, and is given decision rule ρ with probability $p_{RULE}(\rho|\eta, T)$. p_{RULE} is always discrete (as datasets are necessarily finite), is uniform across the predictors, and uniform across each predictor's observed range. We will reference this as the CGM prior.

F.2 Our Model

$$
\begin{aligned}
y_i &= f(\boldsymbol{x}_i, z_i) + \epsilon_i = \mu(\boldsymbol{x}_i) + \tau(\boldsymbol{x}_i) z_i + \epsilon_i \\
\mu(\boldsymbol{x}_i) &\sim BART(\boldsymbol{x}_i; 200\, trees,\ \beta = 2,\ \eta = 0.95) \\
\tau(\boldsymbol{x}_i) &\sim BART(\boldsymbol{x}_i; 50\, trees,\ \beta = 3,\ \eta = 0.25)
\end{aligned}
$$

$$OR$$

$$
\begin{aligned}
y_i \sim & N(\mu(\boldsymbol{x}_i) + \tau(\boldsymbol{x}_i) z_i, \sigma^2) \\
\mu(\boldsymbol{x}_i) = & \sum_{j=1}^{200} g(\boldsymbol{x}_i; \boldsymbol{\mu}_j, T_{\mu j}) \\
& \mu_{j\ell} \sim N(0, \sigma_\mu^2) \\
& \sigma_\mu^2 \sim C_+(\text{median} = 2{*}\text{SD(Y)}) \\
& T_{\mu j} \sim CGM(\beta = 2,\ \eta = 0.95) \\
\tau(\boldsymbol{x}_i) = & \sum_{k=1}^{50} g(\boldsymbol{x}_i; \boldsymbol{\tau}_k, T_{\tau k}) \\
& \tau_{k\ell} \sim N(0, \sigma_\tau^2) \\
& \sigma_\tau^2 \sim N_+(0, \text{median} = \text{SD(Y)}) \\
& T_{\tau k} \sim CGM(\beta = 3,\ \eta = 0.25) \\
\sigma^2 \sim & p(\sigma^2)
\end{aligned}
$$

F.3 Posterior

$$
\begin{aligned}
p(\cdot|X,Y,Z) &\propto p(y|\cdot)p(\mu|T_\mu)p(T_\mu)p(\tau|T_\tau)p(T_\tau)p(\sigma^2) \\
&= \left[\prod_{i=1}^{n} p(y_i|\mu,\tau,\sigma^2)\right]\left[\prod_{j}^{200}\prod_{\ell\in|T_{\mu j}|} p(\mu_{j\ell}|T_{\mu j})\right]\left[\prod_{j} p(T_{\mu j})\right]\left[\prod_{k}^{50}\prod_{\ell\in|T_{\tau k}|} p(\tau_{k\ell}|T_{\tau k})\right] \\
&\quad * \left[\prod_{k} p(T_{\tau k})\right] p(\sigma^2) \\
&= \prod_{i=1}^{n}\left[(2\pi\sigma^2)^{-1/2} exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu(\boldsymbol{x}_i) - \tau(\boldsymbol{x}_i)z_i)^2\right\}\right] \\
&\quad * \prod_{k}\left[p(T_{\tau k})\prod_{\ell\in|T_{\tau k}|}(2\pi\sigma_\tau^2)^{-1/2} exp\left\{-\frac{1}{2\sigma_\tau^2}(\tau_{k\ell})^2\right\}\right] \\
&\quad * \left[\prod_{j}^{200}\prod_{\ell\in|T_{\mu j}|} p(\mu_{j\ell}|T_{\mu j})\right]\left[\prod_{j} p(T_{\mu j})\right] p(\sigma^2)
\end{aligned}
$$

Then the complete conditional for a single $\tau_{k\ell}$ is

$$
\begin{aligned}
p(\tau_{k\ell}|\cdot) &\propto \prod_{i=1}^{n}\left[(2\pi\sigma^2)^{-1/2}exp\left\{-\frac{1}{2\sigma^2}(y_i-\mu(\boldsymbol{x}_i)-\tau(\boldsymbol{x}_i)z_i)^2\right\}\right] \\
&\quad * \prod_{k}\left[p(T_{\tau k})\prod_{\ell\in|T_{\tau k}|}(2\pi\sigma_\tau^2)^{-1/2}exp\left\{-\frac{1}{2\sigma_\tau^2}(\tau_{k\ell})^2\right\}\right] \\
&\propto \prod_{i\in T_{\tau k_\ell}}\left[exp\left\{-\frac{1}{2\sigma^2}\left(y_i-\mu(\boldsymbol{x}_i)-z_i\sum_{\neq k}g_\tau(\boldsymbol{x}_i;\boldsymbol{\tau}_k,T_{\tau k})-z_i\tau_{k\ell}\right)^2\right\}\right] \\
&\quad * exp\left\{-\frac{1}{2\sigma_\tau^2}(\tau_{k\ell})^2\right\} \\
&\text{Let } r_{ik}=y_i-\mu(\boldsymbol{x}_i)-z_i\sum_{\neq k}g_\tau(\boldsymbol{x}_i;\boldsymbol{\tau}_k,T_{\tau k}) \\
&\propto \prod_{i\in T_{\tau k_\ell}}\left[exp\left\{-\frac{1}{2\sigma^2}\left(r_{ik}-z_i\tau_{k\ell}\right)^2\right\}\right]exp\left\{-\frac{1}{2\sigma_\tau^2}(\tau_{k\ell})^2\right\} \\
&\propto exp\left\{-\frac{1}{2}\left[-2\frac{\sum_{i\in T_{\tau k_\ell}}r_{ik}z_i}{\sigma^2}\tau_{k\ell}+\left(\frac{\sum_{i\in T_{\tau k_\ell}}z_i^2}{\sigma^2}+\frac{1}{\sigma_\tau^2}\right)\tau_{k\ell}^2\right]\right\} \\
&\text{Let } V^{-1}=\left(\frac{\sum_{i\in T_{\tau k_\ell}}z_i^2}{\sigma^2}+\frac{1}{\sigma_\tau^2}\right), \text{ and complete the square:} \\
&\propto exp\left\{-\frac{1}{2V}\left(\tau_{k\ell}-V\frac{\sum_{i\in T_{\tau k_\ell}}r_{ik}z_i}{\sigma^2}\right)^2\right\} \\
\Rightarrow (\tau_{k\ell}|\cdot) &\sim N\left(\frac{V}{\sigma^2}\sum_{i\in T_{\tau k_\ell}}r_{ik}z_i, V\right),\ V=\left(\frac{\sum_{i\in T_{\tau k_\ell}}z_i^2}{\sigma^2}+\frac{1}{\sigma_\tau^2}\right)^{-1}.
\end{aligned}
$$

The conditional distribution for single $\mu_{j\ell}$ is similar, with "$z_i = 1$" and $r_{ij} = y_i - \tau(\boldsymbol{x}_i)z_i - \sum_{\neq j} g(\boldsymbol{x}_i;\boldsymbol{\mu}_j,T_{\mu j})$:

$$
\Rightarrow (\mu_{j\ell}|\cdot) \sim N\left(\frac{V}{\sigma^2}\sum_{i\in T_{\mu j_\ell}}r_{ij}, V\right), \quad V=\left(\frac{|T_{\mu j_\ell}|}{\sigma^2}+\frac{1}{\sigma_\tau^2}\right)^{-1}.
$$

Bibliography

Ang, A. and Bekaert, G. (2007). Stock return predictability: Is it there? *Review of Financial Studies*, 20(3):651–707.

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10(4):1281–1311.

Billio, M., Casarin, R., Ravazzolo, F., and van Dijk, H. K. (2013). Time-varying combinations of predictive densities using nonlinear filtering. *Journal of Econometrics*, 177(2):213 – 232. Dynamic Econometric Modeling and Forecasting.

Bossaerts, P. and Hillion, P. (1999). Implementing statistical criteria to select return forecasting models: what do we learn? *Review of Financial Studies*, 12(2):405–428.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Brennan, M. J., Schwartz, E. S., and Lagnado, R. (1997). Strategic asset allocation. *Journal of Economic Dynamics and Control*, 21(8):1377 – 1403.

Campbell, J. and Shiller, R. (1988). The dividend-price ratio and expectations of future dividends and discount factors. *Review of Financial Studies*, 1(3):195–228.

Campbell, J. Y., Chan, Y. L., and Viceira, L. M. (2003). A multivariate model of strategic asset allocation. *Journal of Financial Economics*, 67(1):41 – 80.

Carriero, A., Clark, T. E., and Marcellino, M. (2016). Large vector autoregressions with stochastic volatility and flexible priors. *Federal Reserve Bank of Cleveland Working Paper*.

Cattaneo, M. D., Crump, R. K., Farrell, M. H., and Schaumburg, E. (2019). Characteristic-sorted portfolios: Estimation and inference. *FRB of NY Staff Report*, 788.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *Ann. Appl. Stat.*, 4(1):266–298.

Chipman, H. A., George, E. I., McCulloch, R. E., and Shively, T. S. (2016). High-dimensional nonparametric monotone function estimation using BART. *arXiv:1612.01619*.

Christoffersen, P. F. and Diebold, F. X. (1998). Cointegration and long-horizon forecasting. *Journal of Business & Economic Statistics*, 16(4):450–458.

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of Finance*, 66(4):1047–1108.

Cochrane, J. H. and Piazzesi, M. (2005). Bond risk premia. *The American Economic Review*, 95(1):138–160.

Dangl, T. and Halling, M. (2012). Predictive regressions with time-varying coefficients. *Journal of Financial Economics*, 106(1):157 – 181.

Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika*, 68(1):265–274.

Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, 20(3):339–350.

Fama, E. F. and Bliss, R. R. (1987). The information in long-maturity forward rates. *The American Economic Review*, 77(4):680–692.

Fama, E. F. and French, K. R. (1993a). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3 – 56.

Fama, E. F. and French, K. R. (1993b). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33(1):3 – 56.

Fama, E. F. and French, K. R. (2008). Dissecting anomalies. *The Journal of Finance*, 63(4):1653–1678.

Fama, E. F. and French, K. R. (2016). Dissecting anomalies with a five-factor model. *The Review of Financial Studies*, 29(1):69–103.

Fama, E. F. and MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, 81(3):607–636.

Fama, E. F. and Schwert, G. (1977). Asset returns and inflation. *Journal of Financial Economics*, 5(2):115 – 146.

Fisher, J. D., Pettenuzzo, D., and Carvalho, C. M. (2019a). Optimal asset allocation with multivariate Bayesian dynamic linear models. *Working Paper*.

Fisher, J. D., Puelz, D. W., and Carvalho, C. M. (2019b). Monotonic effects of characteristics on returns. *Working Paper*.

Freyberger, J., Neuhierl, A., and Weber, M. (2019). Dissecting characteristics nonparametrically. *Review of Financial Studies, forthcoming*.

Gao, X. and Nardari, F. (2018). Do commodities add economic value in asset allocation? New evidence from time-varying moments. *Journal of Financial and Quantitative Analysis*, 53.

Gargano, A., Pettenuzzo, D., and Timmermann, A. G. (2017). Bond return predictability: Economic value and links to the macroeconomy. *Management Science*, 65(2).

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press.

Geweke, J. and Amisano, G. (2011). Optimal prediction pools. *Journal of Econometrics*, 164(1):130 – 141.

Gruber, L. and West, M. (2016). GPU-accelerated Bayesian learning and forecasting in simultaneous graphical dynamic linear models. *Bayesian Analysis*, 11(1):125–149.

Gu, S., Kelly, B. T., and Xiu, D. (2018). Empirical asset pricing via machine learning. *Chicago Booth Research Paper*, 18-04.

Gurkaynak, R. S., Sack, B., and Wright, J. H. (2007). The U.S. Treasury yield curve: 1961 to the present. *Journal of Monetary Economics*, 54(8):2291 – 2304.

Hahn, P. R. and Carvalho, C. M. (2015). Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association*, 110(509):435–448.

Hahn, P. R., Carvalho, C. M., Puelz, D., and He, J. (2018a). Regularization and confounding in linear regression for treatment effect estimation. *Bayesian Analysis*, 13(1):163–182.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. (2018b). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *arXiv:1706.09523v2*.

Harvey, C. R., Liu, Y., and Zhu, H. (2016). ... and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Jegadeesh, N. and Titman, S. (1993). Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1):65–91.

Jegadeesh, N. and Titman, S. (2001). Profitability of momentum strategies: An evaluation of alternative explanations. *The Journal of Finance*, 56(2):699–720.

Johannes, M., Korteweg, A., and Polson, N. (2014). Sequential learning, predictability, and optimal portfolio returns. *The Journal of Finance*, 69(2):611–644.

Kim, C.-J., Morley, J. C., and Nelson, C. R. (2005). The structural break in the equity premium. *Journal of Business & Economic Statistics*, 23(2):181–191.

Koop, G., Korobilis, D., and Pettenuzzo, D. (2018). Bayesian compressed vector autoregressions. *Journal of Econometrics*.

Lettau, M. and Ludvigson, S. (2001). Consumption, aggregate wealth, and expected stock returns. *The Journal of Finance*, 56(3):815–849.

Lettau, M. and Van Nieuwerburgh, S. (2008). Reconciling the return predictability evidence. *The Review of Financial Studies*, 21(4):1607–1652.

Lewellen, J. (2004). Predicting returns with financial ratios. *Journal of Financial Economics*, 74(2):209 – 235.

Ludvigson, S. C. and Ng, S. (2009). Macro factors in bond risk premia. *The Review of Financial Studies*, 22(12):5027–5067.

McCarthy, D. and Jensen, S. T. (2016). Power-weighted densities for time series data. *The Annals of Applied Statistics*, 10(1):305–334.

Pastor, L. and Stambaugh, R. F. (2001). The equity premium and structural breaks. *The Journal of Finance*, 56(4):1207–1239.

Patton, A. J. and Timmermann, A. (2010). Monotonicity in asset returns: New tests with applications to the term structure, the CAPM, and portfolio sorts. *Journal of Financial Economics*, 98(3):605–625.

Paye, B. S. and Timmermann, A. (2006). Instability of return prediction models. *Journal of Empirical Finance*, 13(3):274–315.

Pettenuzzo, D. and Ravazzolo, F. (2016). Optimal portfolio choice under decision-based model combinations. *Journal of Applied Econometrics*, 31(7):1312–1332. jae.2502.

Pettenuzzo, D. and Timmermann, A. (2011). Predictability of stock returns and asset allocation under structural breaks. *Journal of Econometrics*, 164(1):60 – 78. Annals Issue on Forecasting.

Pettenuzzo, D., Timmermann, A., and Valkanov, R. (2014). Forecasting stock returns under economic constraints. *Journal of Financial Economics*, 114(3):517 – 553.

Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. *The Review of Economic Studies*, 72(3):821–852.

Puelz, D., Hahn, P. R., and Carvalho, C. M. (2015). Optimal ETF selection for passive investing. *arXiv:1510.03385*.

Puelz, D., Hahn, P. R., and Carvalho, C. M. (2017a). Regret-based selection for sparse dynamic portfolios. *arXiv:1706.10180*.

Puelz, D., Hahn, P. R., and Carvalho, C. M. (2017b). Variable selection in seemingly unrelated regressions with random predictors. *Bayesian Analysis*, 12(4):969–989.

Puelz, D. W. et al. (2018). *Regularization in econometrics and finance*. PhD thesis, University of Texas at Austin McCombs School of Business.

Rapach, D. E., Strauss, J. K., and Zhou, G. (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *The Review of Financial Studies*, 23(2):821–862.

Shively, T. S., Sager, T. W., and Walker, S. G. (2009). A Bayesian approach to non-parametric monotone function estimation. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 71(1):159–175.

Starling, J. E., Murray, J., Carvalho, C. M., Bukowski, R., and Scott, J. G. (2019). Bart with targeted smoothing: An analysis of patient-specific stillbirth risk. *arXiv:1805.07656*.

Thornton, D. L. and Valente, G. (2012). Out-of-sample predictions of bond excess returns and forward rates: An asset allocation perspective. *Review of Financial Studies*, 25(10):3141 – 3168.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Viceira, L. (1997). Testing for structural change in the predictability of asset returns. *Unpublished*.

Wachter, J. A. and Warusawitharana, M. (2009). Predictable returns and asset allocation: Should a skeptical investor time the market? *Journal of Econometrics*, 148(2):162 – 178.

Welch, I. and Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies*, 21(4):1455–1508.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer Series in Statistics. Springer, 2nd edition.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhao, Z. Y., Xie, M., and West, M. (2016). Dynamic dependence networks: Financial time series forecasting and portfolio decisions. *Applied Stochastic Models in Business and Industry*, 32(3):311–332.

Vita

Jared Dale Fisher was born in North Carolina but grew up in Texas. He first attended Brigham Young University, where he received a Bachelor of Science degree in Statistics with a minor in Mathematics in 2012, followed by a Master of Science degree in Statistics in 2014. Later that year, he began his doctoral studies at the University of Texas at Austin, studying Statistics in the Department of Information, Risk, and Operations Management in the McCombs School of Business.

Permanent address: jared.fisher@utexas.edu

This dissertation was typeset with LaTeX by the author.